ESTIMATION OF GLOTTAL SOURCE FEATURES FROM THE SPECTRAL ENVELOPE OF THE ACOUSTIC SPEECH SIGNAL

A Thesis
Presented to
The Academic Faculty

by

Juan Félix Torres

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
August 2010

ESTIMATION OF GLOTTAL SOURCE FEATURES FROM THE SPECTRAL ENVELOPE OF THE ACOUSTIC SPEECH SIGNAL

Approved by:

Professor Elliot Moore, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Monson H. Hayes
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Chin-Hui Lee
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Kevin Haas
School of Civil and Environmental Engineering
Georgia Institute of Technology

Professor Hongwei Wu
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: 10 May 2010

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Elliot Moore, whose support and guidance were essential to the completion of this work. I am grateful to every member of my committee for taking the time to evaluate my work, provide insightful suggestions, and point out directions for improvement. My progress through the graduate program was greatly facilitated by the diligence of Ms. Marilou Mycko, whom I thank for kindly and patiently answering so many administrative inquiries. Thanks also to my lab colleagues at GT-Savannah for their kind encouragement. I wish the best of luck to all of you.

During my time at Georgia Tech, I've had the good fortune of learning from many passionate, dedicated professors who will surely serve as role models during my future endeavors in academia. I would like to express particular gratitude to Dr. James McClellan, for providing a greatly instructive environment for a then-inexperienced teaching assistant, and again to Dr. Moore, for showing me, by example, how to become an effective lecturer.

I would also like to thank Professor Dan Ellis of Columbia University. It was his wonderful introduction to the field of Speech and Audio Processing that inspired me to consider graduate study. I cannot complete this document without thanking Mr. Philip Mumford, the best Computer Science teacher that any high-school kid could hope to have. Thank you, sir, for giving me a head start.

On a personal level, I am deeply grateful to my parents. Through this long and challenging process, I drew strength from my mother: a kind, generous, hard-working woman with a talent for persevering through any obstacle. Thanks mom, literally, for everything.

From my father, a man whose ability to synthesize knowledge continues to impress me, I have learned many things. Among them is the important virtue of questioning every premise and convention. Thanks dad, for encouraging an early interest in taking things apart.

Finally, there are several people whose unwavering support were crucial to the culmination of this process. I wholeheartedly thank Miguel, for innumerable instances of sound advice, Idalia, for lending an empathetic ear, and Josefina, for her eclectic sense of humor. A special acknowledgement goes to Mrs. Sophie Amick, a talented and intuitive yoga teacher, for providing a relaxing space where past and future anxieties can be forgotten and the mind-body-spirit can be experienced without judgement. Namasté.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

CIQ	Closing Quotient
CPIF	Closed-Phase Inverse Filtering
DAP	Discrete All-Pole Modeling
DCT	Discrete Cosine Transform
EGG	Electroglottograph
FMIF	Fu-Murphy Inverse Filtering Algorithm
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
GMR	Gaussian Mixture Regression
GQIF	Glottal Quality Inverse Filtering
GQM	Glottal Quality Measure
HRF	Harmonic Richness Factor
IAIF	Iterative Adaptive Inverse Filtering
IF	Inverse Filtering
LF	Liljencrants-Fant
LP	Linear Prediction
LPA	Linear Predictive Analysis
LTAS	Long-Term Average Spectrum
LTI	Linear Time-Invariant
MFCC	Mel-Frequency Cepstral Coefficient
NAQ	Normalized Amplitude Quotient
OQ	Open Quotient
PLP	Perceptual Linear Prediction
RAPT	Robust Algorithm for Pitch Tracking
RB-GQA	Rank-Based Glottal Quality Assessment

SEF Spectral Envelope Feature

SQ Speed Quotient

VTF Vocal Tract Filter

SUMMARY

Speech communication encompasses diverse types of information, including phonetics, affective state, voice quality, and speaker identity. From a speech production standpoint, the acoustic speech signal can be mainly divided into glottal source and vocal tract components, which play distinct roles in rendering the various types of information it contains. Most deployed speech analysis systems, however, do not explicitly represent these two components as distinct entities, as their joint estimation from the acoustic speech signal becomes an ill-defined blind deconvolution problem. Nevertheless, because of the desire to understand glottal behavior and how it relates to perceived voice quality, there has been continued interest in explicitly estimating the glottal component of the speech signal. To this end, several inverse filtering (IF) algorithms have been proposed, but they are unreliable in practice because of the blind formulation of the separation problem. In an effort to develop a method that can bypass the challenging IF process, this thesis proposes a new glottal source information extraction method that relies on supervised machine learning to transform smoothed spectral representations of speech, which are already used in some of the most widely deployed and successful speech analysis applications, into a set of glottal source features. A transformation method based on Gaussian mixture regression (GMR) is presented and compared to current IF methods in terms of feature similarity, reliability, and speaker discrimination capability on a large speech corpus, and potential representations of the spectral envelope of speech are investigated for their ability represent glottal source variation in a predictable manner. The proposed system was found to produce glottal source features that reasonably matched their IF counterparts in many cases, while being less susceptible to spurious errors.

The development of the proposed method entailed a study into the aspects of glottal source information that are already contained within the spectral features commonly used in speech analysis, yielding an objective assessment regarding the expected advantages of explicitly using glottal information extracted from the speech signal via currently available IF methods, versus the alternative of relying on the glottal source information that is implicitly contained in spectral envelope representations.

CHAPTER I

INTRODUCTION

Spoken language expresses much more than phonetic content. Along with phonetic information, listeners naturally and habitually extract from the acoustic speech signal a diverse set of supplementary cues that are crucial for successful human communication. Over several decades, computational speech processing has sought to build automated systems which can recreate – or even surpass – the information extraction and production abilities of human listeners and speakers with respect to various aspects of spoken language. While there have been many successes in the extraction of certain types of information, particularly phonetic content, other aspects of speech communication, such as vocal affect and voice quality, remain difficult to analyze and reproduce.

From a speech production standpoint, the properties of the acoustic speech signal can be mainly divided into two categories: vocal tract effects, which relate to the shape of the oral and nasal cavities at a particular point in time, and glottal source effects, which relate to the airflow pattern across the glottis as the vocal folds abduct and adduct. These two components of speech play distinct roles in rendering the various types of information contained in the speech signal. Phonetic identity, for instance, is in many cases conveyed by formants – a series of resonances that are controlled by the shape of the vocal tract. For other types of speech content, such as voice identity and voice quality, both the vocal tract and glottal components of speech contain perceptually relevant and complementary information [68, 91], while in the case of emotion and affective state, there has been strong interest in investigating specific relationships to glottal source effects [28, 72, 48, 89, 88, 79, 105].

Although the complementary nature of the vocal tract and glottal components of speech has been well established, in practice, these components are often not explicitly analyzed as separate entities. One reason for this is that these components are not observed separately, but in combination, in the acoustic speech signal, and the separation of two unknown signals from a single observed signal remains, even under simplifying assumptions, an ill-defined blind deconvolution problem. Indeed, some of the most widely deployed and successful speech analysis applications, such as speech recognition and speaker identification / verification, have conventionally focused not on a production model, but on a perceptual one, adopting a warped and smoothed spectral representation of the speech signal that is affected by the state of both the vocal tract and the glottis, and from which the desired discriminatory information is extracted via pattern analysis [30, 61, 97, 96, 76]. Even speech coders that use a linear predictive (LP) representation of speech – due to the all-pole LP filter's ability to model vocal tract resonances – work with a residual signal that is by no means intended to specifically represent the glottal content of speech beyond the presence of pitch marks.

Nevertheless, partly because of the apparent limitations of conventional spectral representations in emerging speech applications such as vocal affect analysis, but also because of an interest in understanding glottal behavior and how it relates to perceived voice quality [21, 7, 87], there has been a continued interest in explicitly estimating and parameterizing the glottal component of the acoustic speech signal. To this end, several inverse filtering algorithms have been proposed over the years in an attempt to estimate and remove the vocal tract component of the signal in order to reveal the glottal component. These algorithms are, however, difficult to use in practice. One issue that has precluded their widespread adoption is that the simplifying assumptions under which inverse filtering algorithms operate are often violated in real speech signals. Although recent work has allowed for the relaxation of some of those

assumptions, this has been at the cost of much increased model complexity, which has lead in some cases to convergence problems [43]. Furthermore, the time-domain representation of speech on which these algorithms operate is susceptible to noise and phase degradation, which have limited the utility of these algorithms to laboratory settings in which high-quality audio capture can be performed. Thus, if glottal information is to be used in practical settings, an information extraction method that overcomes some of the difficulties associated with inverse filtering is needed.

The goal of the work presented in this thesis was to propose and evaluate a new glottal information extraction method which may be easily incorporated into conventional feature extraction frameworks that rely on a low-dimensional short-time spectral representation of speech. The proposed system makes use of supervised machine learning techniques to model transformations from a conventional set of spectral speech features into the desired glottal features. Once trained, the system may be used to quickly estimate glottal features using existing spectral feature sets. The development of the proposed method entails a study into the aspects of glottal source information in normal speech that are already contained within the spectral features commonly used in speech analysis. Therefore, this investigation yields an objective assessment regarding the expected advantages of explicitly using glottal information extracted from the acoustic speech signal via currently available blind-deconvolution methods, versus the alternative of relying on the glottal source information that is implicitly contained in conventional spectral envelope representations.

The remainder of this thesis begins with a brief introduction to basic concepts in speech processing (Chapter 2), wherein the source-filter model of speech production is presented and the basic properties of the glottal and vocal tract components of speech are discussed. Chapter 3 continues the presentation of background material with a review of existing methods for measuring glottal airflow and observing vocal fold behavior, which include visual, impedance, and acoustic signal modalities. Emphasis

is given to the discussion of existing inverse filtering techniques that are designed to estimate glottal behavior using the acoustic speech signal alone. The review proceeds with a survey on the use of glottal source information in speech analysis applications and concludes with a discussion of previous work that is relevant to the estimation of glottal source information from conventional features of the speech spectrum.

The machine learning techniques that were used to develop the proposed method of glottal information extraction are presented in Chapter 4, along with the objective measures and methods used to evaluate its performance. The chapter concludes with a description of TIMIT, a large speech corpus of 630 speakers that was used to train and evaluate the proposed method.

The existing inverse filtering techniques that were chosen to obtain glottal source information for this study are defined and discussed in Chapter 5. These techniques range from a "classical" inverse filtering method to a recently proposed method that may be regarded as "state-of-the-art," and were used to extract baseline glottal features for training the proposed system. Chapter 6 describes the spectral envelope features of the acoustic speech signal that were considered as candidate input features for the proposed system, as well as a comprehensive set of glottal waveform features that were estimated. The chapter then proceeds with an account of the feature extraction procedure for the entire TIMIT corpus. As a preliminary step to the development and evaluation of the proposed system, the measurement reliability of the extracted glottal waveform measures was analyzed in order to select the most suitable inverse filtering method for obtaining the glottal waveform features to be used for training.

The proposed method of glottal waveform feature estimation is presented in Chapter 7. The system was trained using each potential spectral envelope feature

set, and the estimated glottal waveform features were compared to their inverse filtering counterparts in terms of similarity, measurement reliability, and speaker discrimination capability. Chapter 8 provides a summary of the conducted research, draws general conclusions from the combined results, and suggests directions for future work.

CHAPTER II

CONCEPTS

2.1 Speech Production

Speech production occurs as air is expelled from the lungs and travels through the glottis and into the oral and nasal cavities, finally exiting the body through the lips and/or nose. At any given time, the characteristics of the emitted acoustic waveform are controlled by the shape of the oral cavity and lips, the proportion of air that escapes through the nasal cavity, and the state of the glottis. Spectrally, the oral and nasal cavities create a series of perceivable resonances and anti-resonances that change according to the position and movement of the articulators (jaw, tongue, teeth, and lips). In many cases, these resonances are sufficient for identifying phonemes. The vocal folds, which surround the glottis, can be relaxed, in which case air flows freely through the glottis, resulting in a turbulent, 'noisy' sound (unvoiced speech). During voiced speech, the vocal folds are placed within close proximity of each other, and the air pressure from the lungs causes them to vibrate, resulting in an airflow pattern with a fundamental frequency equal to the vocal folds' rate of vibration. In addition to determining the fundamental frequency of voiced speech, the properties of the glottal airflow pattern during voiced phonation are responsible for perceivable differences in voice quality, and contribute to the perception of speaker stress, affective state, and identity, as discussed in Section 3.2.

2.2 The Linear Time-Invariant Source-Filter Model

Although it has been established that speech production is a non-linear process [16, 11, 110, 111] in which the state of the oral and nasal cavities can affect the airflow pattern

across the glottis, and vice-versa, the linear source-filter model of speech production relies on the simplifying assumption that the vocal tract and glottal components of speech are independent and separable. Under this model, the time-domain acoustic speech signal $s[n]$ is understood to consist of the convolution of a *glottal waveform* signal (also called the *glottal source* or the *voice source*) $g[n]$, which represents the volume velocity of airflow across the glottis as a function of time, with the impulse response of a vocal tract filter (VTF) that represents the resonances (formants) and anti-resonances produced in the oral and nasal cavities. To represent the continuous changes to the vocal tract spectrum that occur in speech due to the movement of the articulators, the vocal tract filter ought to be time-varying. However, because the rate of movement of the articulators is limited, the general approach is to convert the source-filter representation of speech into a time-invariant model by segmenting $s[n]$ into 10–30 ms frames $s_k[n]$, and using a separate, time-invariant VTF $V_k(z)$ for each frame k. This linear, time-invariant (LTI) representation enables the application of "classical" DSP and linear systems theory to the processing and analysis of the acoustic speech signal.

The LTI speech production model for a short time frame of speech $s_k[n]$ can be stated in the Z domain as

$$S_k(z) = G_k(z)V_k(z)R_k(z), \tag{1}$$

where $G_k(z)$ is the transfer function of $g_k[n]$ and $R_k(z)$ represents the lip radiation filter, which models the coupling of the vocal tract to the surrounding air volume. $R_k(z)$ is usually approximated as a first-order difference $R_k(z) = R(z) = 1 - \alpha z^{-1}$, with $\alpha \approx 1$. For practical simplicity, the glottal waveform and lip-radiation components are sometimes lumped together, giving rise to the concept of the glottal waveform *derivative*, which is expressed in the time-domain as $g'[n] = g[n] - \alpha g[n-1]$ or in the Z domain as $G_d(z) = G(z)R(z)$.

Figure 1 illustrates the glottal source and vocal tract components of speech for

a 25 ms segment of a vowel utterance. The fundamental period $T_0 = 1/f_0$ of the glottal waveform (Figure 1(a)) determines the fundamental frequency f_0 or *pitch*[1] of the speech signal $s[n]$. The fundamental period of $s[n]$ can be clearly observed from Figure 1(e) and is also apparent in the frequency domain (Figure 1(f)) as the distance between the *pitch harmonics* Hn of $S(e^{jw})$. In addition to a harmonic structure, the glottal waveform derivative spectrum $G_d(e^{jw})$ (Figure 1(b)) generally possesses a bandpass spectral envelope, with a zero at DC and a peak somewhere between $H1$ and $H2$ called the *glottal formant*. At higher frequencies, the main feature of an ideal glottal waveform is its *spectral roll-off*, which is typically in the range -6 dB to -12 dB per octave.

The contribution of the vocal tract filter $V(z)$ (Figure 1(d)) is a set of broadband resonances, or *formants*, which provide perceptual cues that are necessary to identify a phoneme. It can be seen from Figure 1(f) that the broad *envelope* of the resulting speech spectrum is affected by both the formant structure of $V(z)$ and the spectral envelope of $G(z)$.

2.3 The Glottal Cycle

For non-pathological, modal phonation, a typical glottal cycle can be divided into five overlapping phases, as illustrated in Figure 2. The opening phase is generally defined as the portion of the glottal cycle where the vocal folds abduct and there is an increase in airflow through the glottis, up to the instant of maximum abduction, t_p. The closing phase starts at t_p and ends at the glottal closure instant (GCI) t_e, where the vocal folds quickly adduct, resulting in a negative peak in $g'[n]$ (Figure 2(a)). The return phase is the interval starting from t_e, where vocal fold adduction continues until $g'[n]$ is effectively zero. The open phase interval $[0, te]$ spans the opening and

[1]Strictly speaking, pitch is a perceptual phenomenon related to how the human auditory system interprets the harmonic structure of sound. The term will be used here to simply refer to the fundamental frequency of the signal.

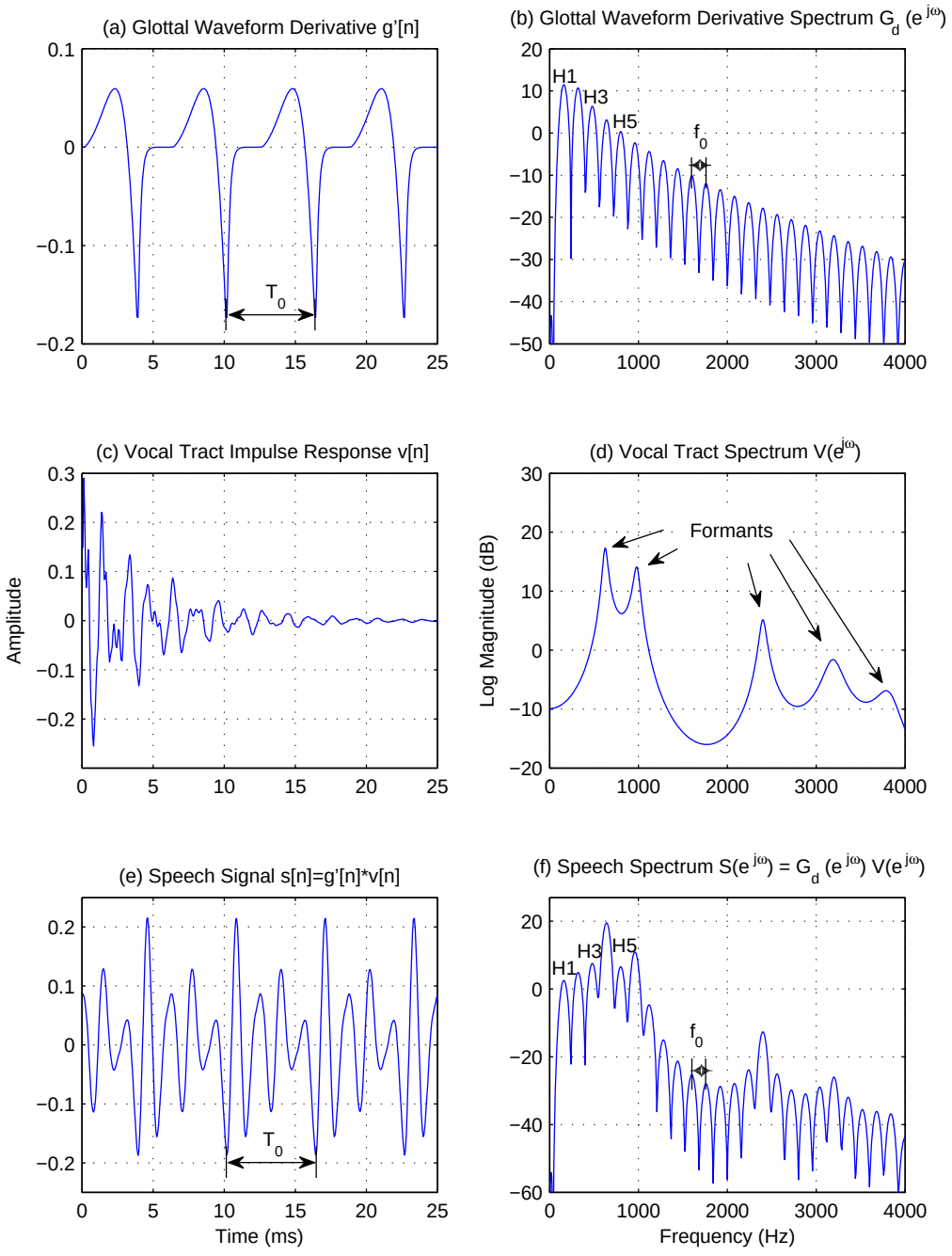


Figure 1: Glottal and vocal tract components of speech, according to the linear time-invariant source-tract speech production model.

closing phases, while the closed phase starts at the GCI (t_e) and continues through the end of the glottal cycle. From a glottal waveform estimate $\hat{g}_k[n]$ or its first-order difference $\hat{g}'_k[n] = \hat{g}_k[n] - \alpha\hat{g}_k[n-1]$, salient features can be computed either by direct measurement or by fitting a parametric model. These features include a set of quotients that measure the relative lengths of the various phases of the glottal cycle, as well as various frequency-domain measures that quantify the slope of spectral roll-off and the relative strength of the lower harmonics. It is these features that are then input to a machine learning algorithm in speech analysis applications that make use of glottal information. Likewise, in formant-based speech synthesis, a set of these features are used as model parameters to generate an artificial glottal waveform. Thus, for these applications, it is the features of the glottal waveform, and not the waveform itself, that are of interest.

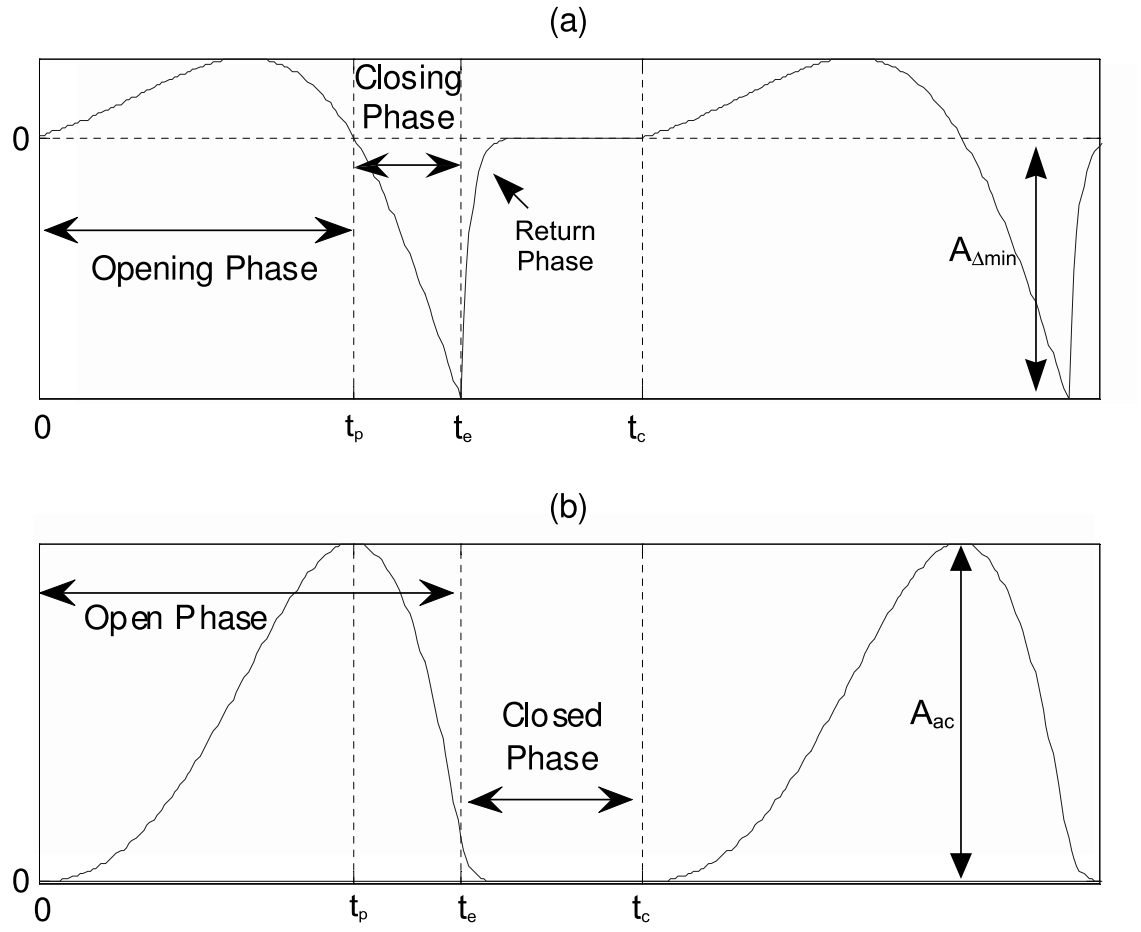


Figure 2: Synthetic (Liljencrants-Fant Model) glottal waveform derivative (a) and corresponding glottal waveform (b).

CHAPTER III

BACKGROUND

3.1 Observation of Glottal Behavior

There exist several methods for directly and indirectly assessing glottal and vocal fold behavior, including direct visual observation of the larynx, measurements of laryngeal impedance, airflow volume velocity measurements at the mouth, and finally, manipulation of the acoustic (pressure) speech signal via inverse filtering. These methods range from the highly-invasive (visual observation) to the minimally invasive (impedance) and the non-invasive (acoustic speech). Their level of invasiveness, combined with the particular signal modality that is observed (image, impedance, airflow), circumscribes the scenarios under which each method can provide useful information. The following sections describe the main issues associated with each type of procedure.

3.1.1 Laryngeal Imaging

Vocal fold motion may be directly observed via several high-speed imaging techniques, such as videokymography [107], stroboscopy [100], and high-speed digital video recording [63], the latter of which allows direct observation of the vocal folds by recording video at 2000–4000 frames per second. In a modern stroboscopic setup, a flexible endoscope is introduced into one nostril and brought close to the glottis, affording an unobstructed view of the vocal folds. However, high-resolution, high-speed video recording typically requires the use of a rigid endoscope and oral access, thus impeding the observation of speech segments where the mouth is not sufficiently open. While immensely useful for the diagnosis and treatment of speech disorders by a speech pathologist, the use of video signals for speech analysis applications is made

difficult by the complex relationship between the image of the vocal folds and the acoustic speech waveform, as the video signal generally contains much more information than the acoustic speech signal, and many details apparent in the video recording may not have significant acoustic effects. The quantitative analysis of high-speed vocal fold video remains the subject of current research [50, 122, 123]. In addition, the invasive nature of direct vocal fold imaging methods not only makes recording large amounts of speech impractical, but can also affect the acoustic speech output due to physical (invasion of the nasal or oral cavity) and psychological (speaker stress and discomfort) reasons. Such unintended modification of the acoustic signal is of particular concern in studies involving voice quality and vocal affect.

3.1.2 Electroglottography

The laryngograph, also known as the electroglottograph or EGG [20], is an external device that measures the impedance across the larynx. This impedance is closely related to the contact area between the vocal folds. The main advantage of the EGG over other methods is that, because its sensors consist of a pair of electrodes placed over the neck, it is able to provide information about vocal fold motion without being affected by the state of the vocal tract and with minimal discomfort to the subject.

Although it is not a measure of airflow through the glottis, there exists a relationship between the EGG signal and the glottal waveform. Specifically, the EGG signal and the glottal waveform are synchronous (have the same fundamental frequency). Furthermore, the time derivative of the EGG signal can be used to approximate the instants of glottal closure and, to a lesser extent, glottal opening[1] [117, 71, 60, 113]. However, due to the fact that the EGG waveform represents the degree of contact between the folds, the EGG provides virtually no information about vocal fold dynamics during the open glottal phase. Conversely, during the closed glottal phase (in

[1]Strictly speaking, EGG-estimated opening and closing instants should be called instants of vocal fold *contacting* and *decontacting*, respectively.

cases where complete glottal closure does occur), additional changes in the contact area between the folds that occur due to the three-dimensional nature of the folds may be registered in the EGG signal even though no changes in airflow take place [15].

Such limitations have resulted in the EGG signal being used primarily for voicing and pitch detection. Nevertheless, due to the minimal invasiveness of the EGG, and because the EGG signal is unaffected by variations in the vocal tract, EGG-derived features were initially identified as potentially desirable targets for the feature transformation method presented in this thesis. The use of EGG information is this manner is subject to the existence of an approximate correspondence between glottal contact and airflow information, since variations in airflow are what lead to acoustic (and perceptual) changes. Appendix B discusses the results of an initial study on the relationship between salient features of the EGG signal and those of the glottal airflow waveform, where features from the two signals were found to be related to each other in an inconsistent and speaker-dependent manner.

3.1.3 Inverse Filtering

Under the assumptions of the LTI source-filter model of speech production, discussed in Section 2.2, it should be possible to compute the glottal waveform signal from the acoustic speech signal if the VTF is known. In practice, when only the acoustic speech signal is observed, the computation of the glottal waveform estimate $\hat{g}_k[n]$ relies on how well the vocal tract filter can be estimated from $s_k[n]$, which is in itself a difficult task because both glottal and vocal tract characteristics influence the observed speech signal, thus turning the joint estimation of the VTF and glottal waveform into a blind deconvolution problem [32, 52, 121]. Nevertheless, assuming the existence of a VTF estimate $\hat{V}_k(z)$ that is valid for time frame k, the glottal waveform estimate $\hat{g}_k[n]$ can

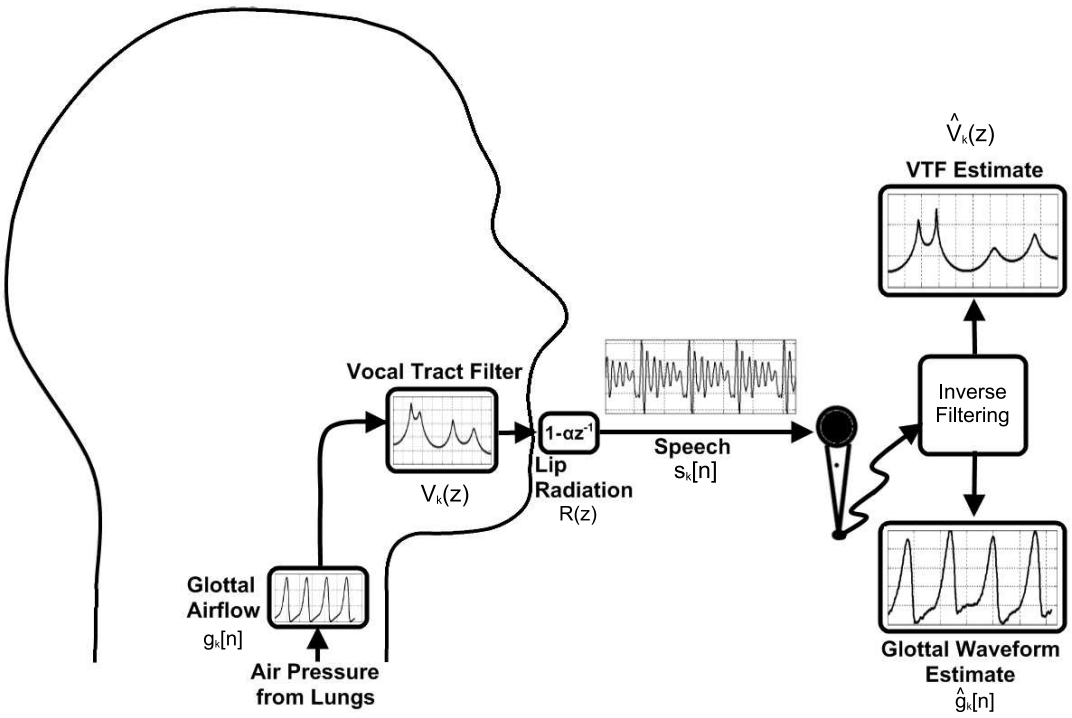


Figure 3: Speech production and analysis according to the LTI source-filter model.

be computed via *inverse filtering* (Figure 3) as follows:

$$\hat{G}_k(z) = \frac{S_k(z)}{\hat{V}_k(z)R(z)}. \tag{2}$$

It should be noted that because of the $R(z)$ term in the denominator of Equation 2, $g_k[n]$ is defined up to an arbitrary constant term, so that it is generally not possible to determine from $g_k[n]$ whether the closed phase represents complete or partial glottal closure[2].

Over time, proposed automatic inverse filtering algorithms have attempted to produce better estimates of the glottal waveform by exploiting various assumptions about the speech production process. What follows is a brief discussion of the main ideas put forth in the inverse filtering literature. The purpose of the following sections is to highlight notable algorithms that are representative of the main thrusts in inverse

[2]The possibility of obtaining an absolute measurement of airflow at the lips using a vented pneumotachograph mask [98, 14] does exist. However, this method is somewhat intrusive as it involves the use of a mask that completely covers the nose and mouth.

filtering research, with emphasis on recent work, and is not meant as an exhaustive review. A recent, in-depth survey on inverse filtering may be found in [120], while [29] gives a thorough review of the various glottal source models that are sometimes incorporated into inverse filtering algorithms. The specific algorithms that were used in this thesis are described in detail in Chapter 5.

3.1.3.1 Closed-Phase Analysis

Under the assumptions of the source-filter model, it has been widely accepted that the best interval of the glottal cycle for estimating the glottal and vocal tract components of speech begins at the instant of glottal closure [121, 114]. The primary rationale behind closed-phase analysis is the approximate lack of vocal tract excitation during the closed glottal phase, where the air from the sub-glottal region is generally cut-off or greatly reduced. This implies a corresponding reduction in the exchange of energy between the sub-glottal (i.e., air from the lungs) and the supra-glottal (i.e., vocal tract) regions, which maximizes the validity of the independent source-filter assumption. Based on these assumptions about the closed glottal phase, closed-phase inverse filtering is carried out by aligning an analysis window much shorter than the length of the pitch cycle with the instant of glottal closure and estimating the VTF using the covariance method of linear prediction (LP) analysis [121].

In practice, however, glottal closure is not always complete, and even when the vocal folds do close completely, the closed phase is often not long enough to contain the entire analysis window. The problem of short closed phases can be somewhat ameliorated by the use of multiple pitch-synchronous estimation windows [18, 91], which allows individual windows to be smaller. However, the inability to handle incomplete closures remains a major issue with the closed-phase approach. In addition, it has been observed that even small estimation errors of the location of the closed

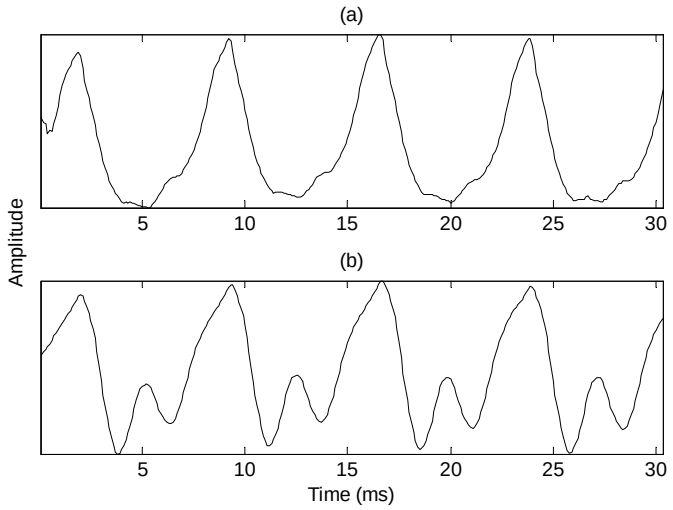


Figure 4: Glottal waveform estimates from the same speech segment, obtained using a small linear prediction analysis window centered inside (a) the closed glottal phase and (b) the open glottal phase. The estimate in (b) is corrupted by first-formant ripple.

glottal phase can result in substantially different vocal tract estimates [114]. Incorrect estimation of the vocal tract can lead to, for example, glottal estimates that are corrupted by first-formant ripple, as shown in Figure 4. Such errors can adversely affect the computation of glottal waveform features.

3.1.3.2 Incorporation of Parametric Glottal Waveform Models

An alternative approach to closed-phase analysis is to begin the inverse filtering process by assuming a parametric model of the glottal waveform and preliminarily removing its effects from the speech signal. The vocal tract filter may then be estimated from the pre-processed signal using data from either the entire pitch cycle [6], or an extended "pseudo-closed phase" [4], and progressive refinement of the $V_k(z)$ and $g_k[n]$ can occur over a number of iterations. These algorithms usually rely on a low-complexity model of the glottal source to produce the initial vocal tract estimate, and may therefore suffer in cases where the actual glottal waveform does not fit the assumed model. A variation of this approach [43] assumes the more complex

Liljencrants-Fant (LF) model of the glottal waveform [41] from the start of the procedure. While their approach was shown to perform well on synthesized speech, the authors noted that in the case of real speech, "problems still exist with regard to robustness".

More recently, an algorithm that jointly estimates LF-model and vocal-tract parameters according to a time-varying autoregressive model was introduced [44]. The main advantage of this algorithm is that it allows the VTF to change at every sample. The use of a time-varying VTF within the speech frame allows the algorithm to represent source-tract interaction by incorporating its effects into a VTF that is allowed to vary across the various phases of the glottal cycle. The authors point out that, as with other approaches that make use of parametric glottal models, "assuming an exact initialization and a converged optimization, the accuracy of the proposed method is finally determined by the ability of the LF model to model phonation".

3.1.3.3 Zeros of Z-transform Representation

A recent, unconventional approach to glottal waveform estimation [17] relies on precisely centering two pitch periods inside a Blackman window. If this is achieved precisely, the glottal waveform can be considered mostly anti-causal and the vocal tract response mostly causal. Upon taking the Z-transform of the entire frame, examination of the zeros of the Z-domain polynomial reveals that the source spectrum contains mostly zeros outside the unit circle. By reconstructing the time-domain signal using only these zeros, an approximation of the glottal waveform can be obtained. Pending issues with this algorithm involve the requirement to precisely segment and align two pitch periods, as well as the inability to separate the glottal source zeros corresponding to the 'causal' portion of the glottal waveform (i.e. inside the unit circle) from the zeros of the vocal tract. In addition, the assumption of glottal waveform causality and anti-causality is based on experiments with synthetic waveforms

generated by the LF model. Although [104] showed the algorithm working well on a short segment of real speech, its adequacy for use on real speech data has yet to be evaluated.

3.1.3.4 Summary

Existing inverse filtering approaches share a common set of limitations, the main issue being the ambiguity in determining which properties of the speech frame belong to $V_k(z)$ and which belong to $g_k[n]$. The various assumptions that are made about each of these signals in an effort to tackle this blind deconvolution problem has lead to algorithms that may not perform well when these assumptions are not met in real speech data. On the other hand, the relaxation of these assumptions often leads to more complex models and algorithms that are susceptible to convergence problems.

3.2 Use of Glottal Features in Speech Analysis

Despite the limitations of inverse filtering procedures, many studies have shown that features related to the glottal source are an important component for the analysis of various forms of non-phonetic content in speech. What follows is a brief review of this work, focusing on recent results related to voice identity and the affective component of speech. While the accuracy of glottal waveform estimation is often mentioned as a limiting factor, these studies do show benefits from incorporating glottal source information into the analysis.

Glottal source features derived from glottal waveforms obtained via inverse filtering have been shown to somewhat complement mel-frequency cepstral coefficients (MFCCs), a representation of the spectral envelope of $s_k[n]$, in speaker identification applications. Recent work [51] shows that the combination of the glottal waveform's cepstral coefficients with conventional MFCCs can reduce the speaker misclassification rate by over 1% on the TIMIT corpus and over 3% on a corpus with a smaller number

of speakers. These results are consistent with previous studies in which glottal waveform features obtained via parametric model fitting were found improve classification rate when combined with vocal tract and prosodic features [18], and MFCCs [91] in a speaker ID scenario. Other studies have obtained similar results by extracting glottal source features using novel representations of the linear-prediction (LP) residual. In these studies, the LP filter was computed over the entire speech frame and no explicit effort was made to obtain an accurate estimate of the glottal waveform. While this linear prediction procedure does not amount to inverse filtering, as the LP filter will likely represent some of the spectral properties of the glottal source (such as spectral tilt and the glottal formant), and vice-versa, the information contained in the LP residual can be regarded as being mainly glottal-source specific, while the LP filter represents most of the properties of the vocal tract. Wavelet coefficients [127] and features derived from the phase spectrum [83] of the LP residual have each been found to reduce the classification error of an MFCC-based speaker identification system by a small percentage.

Perceptual evidence of the complementarity between glottal source and vocal tract information in the formation of voice identity was obtained from a voice conversion study [68] where the voice quality of a set of source speakers was converted to that of a set of target speakers by performing a statistical transformation of the source speakers' LP filters so that they resembled the LP filters of the target speakers. Perceptual evaluation revealed that in some cases, listeners judged the converted speech as coming from a "third" speaker having characteristics from both the source and target speaker. Improved perceptual results were obtained by also modifying the LP residual during conversion [67]. A more recent voice conversion study [31] obtained positive results by performing inverse filtering and transforming both the VTF and glottal waveform components of the speech signal.

The glottal source has also been shown to play an important role in the recognition

of stress and affective states in speech. Several studies have established the relationship between specific changes in glottal source features and laryngeal voice quality [21, 7, 87]. Furthermore, a perceptual study on the relationship between voice quality and the communication of emotion [48] included glottal waveform parameters in the set of variables that were modified to synthesize speech stimuli with varying voice quality. The results have suggested that voice quality plays an important role in the differentiation of subtle emotions, although the perceptual mappings between voice quality and emotion labels varied across listeners.

Direct links between emotional speech and glottal waveform parameters have been studied as well. Glottal waveform features have been shown to vary significantly across utterances representing different types of simulated (acted) emotion [72, 3, 119] as well as induced and simulated stress [28, 116, 54], although the observation has been made that the acoustic expression of each emotion varies significantly across speakers [72] and genders [3]. In addition, glottal features have been shown to be useful for the classification of clinically-depressed subjects from their speech [88, 78], and have been found to perform better than vocal tract features in this task when combined with prosodic features [79].

Other studies have established glottal waveform features as helpful in extracting emotional information from the speech signal beyond what may be obtained from conventional prosodic features (pitch and energy). A feature selection and classification experiment on acted emotional speech [42] found features related to the length and symmetry of the open glottal phase to be the highest ranked among a large set of glottal, loudness, and pitch features. Similarly, a recent study [105] found glottal waveform features to be useful in differentiating between pairs of emotions with similar pitch statistics, while a study on deceptive speech found a similar feature set to generally perform better than pitch in detecting the stress arising from deception [115]. These results are in agreement with the study described in [86], where MFCCs

computed over a limited frequency range of 20–300 Hz (MFFC-low) were found to be similarly useful to full-band MFCCs and more useful than pitch features for emotion classification. Over a frequency range of 20–300 Hz, the MFCC-low features are less influenced by the vocal tract, since the first formant is near 300 Hz only for high vowels (e.g., /iy/, /ux/), and higher formants are out of range [95]. The spectral characteristics of the glottal waveform, on the other hand, are mostly within range, as the glottal formant is usually between the first two pitch harmonics. Thus, particularly for male speakers, the MFCC-low features are not only representing pitch information, but also contain glottal source information, which may explain their additional ability to discriminate emotions in comparison to the pitch feature.

3.3 Related Work

The difficulties related to inverse filtering (Section 3.1.3) have motivated the development of analysis methods that can obtain at least some amount of information about glottal airflow directly from the acoustic speech signal without the need to explicitly estimate the glottal waveform. The long-term average spectrum (LTAS) [74] is a classical method that provides insight into the glottal source spectrum directly from the speech signal. The LTAS is computed over a long sample of speech (>30 seconds) and operates under the assumption that the resonances of a time-varying vocal tract will average out across the long sample into an approximately flat spectrum, revealing an average spectral estimate of the underlying glottal source. A major limitation of the LTAS procedure is the loss of time resolution, which impedes the use of this approach in situations where the object of the analysis may be of short duration (e.g. vocal affect or voice quality analysis) or where a long speech sample is not available. In addition, the results obtained via this method may be undesirably affected by both the (time-invariant) transfer function of the recording channel and by the choice of speech sample, as the assumption of a flat long-time average vocal tract spectrum

critically depends on the balance of phonetic content within the sample.

An alternative approach [55, 65] allows limited estimation of glottal source harmonics at the speech frame level (≈20 ms) by applying a correction formula to the speech spectrum in order to remove the influence of the vocal tract formants, thus revealing an estimate of the first two glottal source harmonics ($H1$, $H2$). This approach can be regarded as a "partial" inverse filtering in the spectral domain, and is useful when the phase of the speech signal has been corrupted by the recording equipment (thus hindering the use of time-domain inverse filtering). However, the success of this procedure is still dependent on an accurate estimate of $V_k(z)$.

The proposed idea of transforming conventional, frame-level *spectral envelope features* (SEFs) into glottal waveform features is motivated by studies that have shown, either explicitly or implicitly, a relationship between the spectral envelope of speech and characteristics of the glottal waveform. It has long been recognized that spectral envelope features reflect a combination of the vocal tract and glottal contributions to the speech signal. In [66], favorable results were obtained in speech recognition experiments through the use of a liftering procedure designed to remove the influence of glottal flow characteristics from cepstral features. Similarly, the authors in [75] used, for the purpose of improving the matching of real speech to a stored codebook in an articulatory speech coding application, a set of optimal lifters to minimize the influence of the glottal flow on the cepstral coefficients. The lifters were derived by examining the effects of variations in the parameters of a physical vocal fold model on the cepstrum of the speech signal generated by an articulatory synthesizer. Conversely, SEFs which were initially designed to capture the phonetic content of speech, have shown significant discrimination ability in certain speech analysis tasks where the glottal source is believed to play a major role. For example, the work in [124] classified creaky phonation, largely a glottal effect, using a representation of the spectral magnitude envelope of speech, while the authors in [34] identified pathological

phonation using a combination of spectral envelope features and pitch information.

An explicit relationship between the glottal waveform and the speech spectrum has been established for synthesized speech in [40, 12, 36], where the magnitude spectrum of time-domain parametric models of the glottal waveform is analytically derived. These studies demonstrate how specific variations of the time-domain parameters lead to systematic changes in the magnitude spectrum of the synthesized glottal waveforms, therefore affecting the spectrum of the speech output. Specifically, variations in the relative duration of the changes in open phase, and the degree of asymmetry between the opening and closing phases have been shown to be related, respectively, to the center frequency and bandwidth of the *glottal formant*, which in turn affects the shape of the speech signal's spectral envelope. In addition, the extent of the return phases has been shown alter the spectral roll-off of the speech spectrum.

On real speech signals, recent work [76] has explored the estimation of fundamental frequency (f_0), perhaps the most easily measurable glottal effect, from SEFs, obtaining high accuracy via a hybrid HMM-GMM (Hidden Markov Model - Gaussian Mixture Model) system. In their approach, the HMMs are used to classify the speech into monophones, and an associated GMM is then used to transform MFCCs into f_0 estimates. Interestingly, it was found that an accuracy level near that achieved by the HMM-GMM system was obtained using a single, larger GMM.

Collectively, these studies suggest that the spectral envelope of the acoustic speech signal may contain enough separable information about the characteristics of the glottal waveform as to allow for a direct transformation from the speech spectral envelope into glottal waveform features, thus bypassing explicit glottal waveform estimation and the challenging process of inverse filtering.

CHAPTER IV

EVALUATION METHODS, TOOLS, AND DATA

4.1 Introduction

Given a set of spectral envelope features and a corresponding set of glottal waveform features, such as those obtained via inverse filtering, the development and evaluation of a system that can transform the former into the latter requires a model that can learn and perform the mapping across feature spaces, as well as a set of measures to evaluate the accuracy of the transformation. Section 4.2 presents a statistical model and training method capable of learning an arbitrary multivariate distribution from a multidimensional dataset. The learned distribution can then be used to perform a regression from one set of variables into another, as described in Section 4.3. Objective measures that will be used to evaluate both the measurement reliability of the glottal waveform features as well as the accuracy of the feature transformation are described in Section 4.4. Finally, a procedure for evaluating the speaker separation ability of a given set of features is described in Section 4.5, and the dataset to be used in this study is described in Section 4.6.

4.2 Statistical Feature Modeling with Gaussian Mixtures

A Gaussian mixture model represents the multivariate probability distribution function of a feature vector $\mathbf{w}$ as

$$f(\mathbf{w}) = \sum_{i=1}^{N} \pi_i \, \mathcal{N}(\, \mathbf{w};\, \mu_i, \mathbf{\Sigma}_i \,), \tag{3}$$

where $\sum_{i=1}^{N} \pi_i = 1$ and $\mathcal{N}(\, \mathbf{w};\, \mu, \mathbf{\Sigma} \,)$ is a multivariate Gaussian distribution with mean vector μ and covariance matrix $\mathbf{\Sigma}$. Given a set of observation vectors $\mathbf{W} = [\mathbf{w}_1 \mathbf{w}_2 ... \mathbf{w}_M]$, the maximum-likelihood parameters $\{\pi_i, \mu_i, \mathbf{\Sigma}_i\}$ for the N Gaussian

mixtures can be estimated via the Expectation Maximization (EM) algorithm [112, 49], which iteratively improves the log-likelihood function

$$\phi_f(\mathbf{W}) = \frac{1}{M}\sum_{m=1}^{M}\log(f(\mathbf{w}_m)). \tag{4}$$

The EM algorithm needs to be given initial estimates of the GMM parameters, which may be obtained by a clustering procedure such as k-means. Because the EM algorithm converges to a local and not a global maximum, in practice it is often run several times with different initializations, and the model from the best run is retained.

The number of Gaussian mixtures N controls the complexity of the model, and it can be shown that given a sufficiently large N, a GMM can represent any arbitrary distribution function. In practice, larger values of N increase the number of parameters, thus requiring more observations for training and increased computation time[1].

4.3 Feature Transformation via GMM Regression

If an observation $\mathbf{w} = \left[\mathbf{x}^T\,\mathbf{y}^T\right]^T$ is said to consist of the concatenation of a source feature vector $\mathbf{x}$ (e.g. spectral envelope features) and a target feature vector $\mathbf{y}$ (e.g. glottal waveform features), then, assuming that $\mathbf{w}$ is distributed according to the GMM of Equation 3, the nonlinear *Gaussian mixture regression* (GMR) function that minimizes the mean-squared error between the actual $\mathbf{y}$ and its estimate $\hat{\mathbf{y}}$ is given as [67, 106, 82]

$$\begin{aligned}\hat{\mathbf{y}} &= \mathcal{F}_f(\mathbf{x}) = E(\mathbf{y}|\mathbf{x}) \\ &= \frac{\sum_{i=1}^{N}\pi_i\,\mathcal{N}(\,\mathbf{x};\mu_{i\mathrm{X}},\mathbf{\Sigma}_{i\mathrm{XX}}\,)\left[\mu_{i\mathrm{Y}}+\mathbf{\Sigma}_{i\mathrm{YX}}\mathbf{\Sigma}_{i\mathrm{XX}}^{-1}(\mathbf{x}-\mu_{i\mathrm{X}})\right]}{\sum_{j=1}^{N}\pi_j\,\mathcal{N}(\,\mathbf{x};\mu_{j\mathrm{X}},\mathbf{\Sigma}_{j\mathrm{XX}}\,)},\end{aligned} \tag{5}$$

where

$$\mu_i = \begin{bmatrix}\mu_{i\mathrm{X}}\\ \mu_{i\mathrm{Y}}\end{bmatrix},\quad \mathbf{\Sigma}_i = \begin{bmatrix}\mathbf{\Sigma}_{i\mathrm{XX}} & \mathbf{\Sigma}_{i\mathrm{XY}}\\ \mathbf{\Sigma}_{i\mathrm{YX}} & \mathbf{\Sigma}_{i\mathrm{YY}}\end{bmatrix}.$$

[1] An approach which has proved useful in speaker identification applications [96] is to increase N and to decrease the number of model parameters by restricting $\mathbf{\Sigma}_i$ to be diagonal.

Thus, the ability to transform spectral envelope features into glottal features is contingent upon their local covariances $\mathbf{\Sigma}_{i\mathrm{YX}}$. The proposed feature transformation method then consists of training a GMM f with a suitable number of mixtures N to estimate the joint distribution of the spectral and glottal feature observations $[\mathbf{w}_1\mathbf{w}_2...\mathbf{w}_M]$ and constructing the transformation function $\mathcal{F}_f(\mathbf{x})$ according to Equation 5. Once this function has been learned, an unseen set of spectral feature observations $[\mathbf{x'}_1\mathbf{x'}_1...\mathbf{x'}_K]$ can be transformed into glottal feature estimates as $\left[\hat{\mathbf{y}}'_1\hat{\mathbf{y}}'_2...\hat{\mathbf{y}}'_K\right] = [\mathcal{F}_f(\mathbf{x'}_1)\mathcal{F}_f(\mathbf{x'}_2)...\mathcal{F}_f(\mathbf{x'}_K)]$.

4.4 Measures of Similarity between Two Observation Sets

To evaluate and compare the extent to which two estimates of a scalar feature (or two different scalar features) approximate each other, multiple observations are required. Given two sets of K feature observations $u = [u_1, u_2...u_K]$ and $v = [v_1, v_2...v_K]$ the *linear correlation coefficient*

$$r_c(u,v) = \frac{\sum_{k=1}^{K}(u_k - \bar{u})(v_k - \bar{v})}{\sqrt{\sum_{k=1}^{K}(u_k - \bar{u})^2 \sum_{k=1}^{K}(v_k - \bar{v})^2}}, \tag{6}$$

where

$$\bar{u} = \frac{1}{K}\sum_{k=1}^{K} u_k, \quad \bar{v} = \frac{1}{K}\sum_{k=1}^{K} v_k,$$

can be used to measure the similarity between them. The correlation coefficient, valued between -1 and 1, reflects the strength and direction of the linear relationship between u and v, with $r_c = 0$ indicating a lack of *linear* relationship and $r_c < 0$ indicating a tendency for u to increase linearly as v decreases.

Another measure of similarity between u and v is the *coefficient of determination* [33], defined as

$$r_d(u,v) = 1 - \frac{\sum_{k=1}^{K}(u_k - v_k)^2}{\sum_{k=1}^{K}(u_k - \bar{u})^2}. \tag{7}$$

This measure takes a maximum value of 1 if and only if $u = v$, and can be interpreted as a measure of the mean-squared error between u and v relative to the variance of

u:

$$r_d(u,v) = 1 - (\frac{K}{K-1})\frac{MSE(u,v)}{\sigma_u^2}, \tag{8}$$

where

$$MSE(u,v) = \frac{1}{K}\sum_{k=1}^{K}(u_k - v_k)^2, \quad \text{and} \quad \sigma_u^2 = \frac{1}{K-1}\sum_{k=1}^{K}(u - \bar{u})^2$$

For large K, a value of r_d close to zero implies that the mean squared error between u and v is as great as the variance of u.

There are two important differences between these measures. Unlike the correlation coefficient, r_d is not symmetric unless u and v have equal variance ($r_d(u,v) = r_d(v,u) \Leftrightarrow \sigma_u^2 = \sigma_v^2$). In addition, given a perfectly linear relationship $\forall\, k : v_k = \alpha u_k + \beta$, the correlation coefficient $r_c(u,v)$ will be equal to 1.0 while $r_d(u,v)$ will depend on the slope α and offset β of the linear relationship. Thus, r_d decreases with respect to differences of scale and bias between u and v while r_c remains unaffected by such differences. Furthermore, if u and v share the same mean and variance, it can be shown by application of the Cauchy-Schwarz inequality that the square of the correlation coefficient becomes an upper bound for the coefficient of determination (i.e., $r_d <= r_c^2$, with equality if u and v are linearly related).

A third similarity measure, which can be useful for quantifying non-linear relationships between u and v, is the Spearman rank-correlation coefficient r_r, given by

$$r_r(u,v) = \frac{\sum_{k=1}^{K}(\Gamma_{u_k} - \bar{\Gamma}_u)(\Gamma_{v_k} - \bar{\Gamma}_v)}{\sqrt{\sum_{k=1}^{K}(\Gamma_{u_k} - \bar{\Gamma}_u)^2 \sum_{k=1}^{K}(\Gamma_{v_k} - \bar{\Gamma}_v)^2}}, \tag{9}$$

where Γ_{u_k} and Γ_{v_k} denote the ranks for the k^{th} observation of u and v, respectively, and $\bar{\Gamma}_u$, $\bar{\Gamma}_v$ denote the average rank of each variable. From Equations 6 and 9, it can be seen that r_r is computed in exactly the same way as the linear correlation coefficient, except that the values of u and v are replaced by their ranks (i.e., their order within the set of K observations). By basing the similarity measure on ranks, r_r is invariant to any monotonic transformation of u and v, and is therefore able to

measure the strength of an arbitrary monotonic, non-linear relationship between the variables.

4.5 Maximum Likelihood Pairwise Speaker Classification

Another way to determine the usefulness of a scalar feature or a feature vector is to evaluate its performance in a speech analysis application. As this study focuses on estimation of glottal waveform features, which are an important component of voice quality and voice identity (as discussed in Section 3.2), it is useful to compare different features and estimation methods by their ability to correctly distinguish between the voices of different speakers. Here, the goal is not to improve upon the already high performance of current speaker identification systems, but to use the speaker ID system as a way to gather information about the discrimination ability of the features, with the assumption that noisy or poorly estimated features will show a decreased ability to correctly distinguish between the voices of different-speaker pairs.

As it is already widely used in speaker identification [97, 96, 91, 39, 77, 51] a GMM classifier will be adopted here. In the case of binary classification of speaker pairs, two separate GMMs, f_{ω_1} and f_{ω_2} are trained with the EM algorithm as described in Section 4.2, using the training observations $[\mathbf{z}_{\omega_1,1}\ \mathbf{z}_{\omega_1,2}\ ...\ \mathbf{z}_{\omega_1,M}]$, from speaker ω_1 and $[\mathbf{z}_{\omega_2,1}\ \mathbf{z}_{\omega_2,2}\ ...\ \mathbf{z}_{\omega_2,M}]$ from speaker ω_2, respectively. Then, for the k^{th} observation $\hat{\mathbf{z}}_k$ from an independent test data set, the most probable speaker $\hat{c}_k \in \{\omega_1, \omega_2\}$ is chosen according to the Bayes classification rule [112], given as

$$P(\omega_1|\,\hat{\mathbf{z}}_k) \underset{\hat{c}_k=\omega_2}{\overset{\hat{c}_k=\omega_1}{\gtrless}} P(\omega_2|\,\hat{\mathbf{z}}_k), \tag{10}$$

which can be rewritten in terms of the class-conditional likelihood functions $p(\mathbf{z}|\,\omega_1) = f_{\omega_1}(\mathbf{z})$ and $p(\mathbf{z}|\,\omega_2) = f_{\omega_2}(\mathbf{z})$ as

$$f_{\omega_1}(\hat{\mathbf{z}}_k)P(\omega_1) \underset{\hat{c}_k=\omega_2}{\overset{\hat{c}_k=\omega_1}{\gtrless}} f_{\omega_2}(\hat{\mathbf{z}}_k)P(\omega_2). \tag{11}$$

Assuming equal *a priori* probabilities for each speaker, $P(\omega_1) = P(\omega_2)$, the decision

rule is then based on selecting the speaker whose GMM outputs the highest likelihood for the input vector:

$$f_{\omega_1}(\hat{\mathbf{z}}_k) \underset{\hat{c}_k=\omega_2}{\overset{\hat{c}_k=\omega_1}{\gtrless}} f_{\omega_2}(\hat{\mathbf{z}}_k). \tag{12}$$

Given a set of actual speaker labels $[c_1 c_2 ... c_K]$, $c_k \in \{\omega_1, \omega_2\}$, and the corresponding classification results $[\hat{c}_1 \hat{c}_2 ... \hat{c}_K]$ obtained according to Equation 12, the *classification rate* can then be computed as

$$C_{rate} = \frac{1}{K} \sum_{k=1}^{K} \phi(c_k, \hat{c}_k), \tag{13}$$

where

$$\phi(c, \hat{c}) = \begin{cases} 1 & \text{if } c = \hat{c} \\ 0 & \text{otherwise.} \end{cases}$$

4.6 Speech Corpus

Evaluation of the measurement reliability, predictability, and speaker discrimination ability of the selected glottal waveform features was performed using a large set of speakers from the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [45, 46]. TIMIT contains a total of 6300 utterances, with 10 sentences spoken by each of 630 native speakers representing 8 major dialects of American English (438 males, 192 females). The 10 sentences contain approximately 30 seconds of speech material per speaker. In total, the corpus contains roughly 5 hours of speech. All speakers were judged by a professional speech pathologist as having no clinical speech pathologies. The 630 corpus sentences are divided into 3 categories: dialect sentences (SA), designed to expose dialectal differences in pronunciation among speakers; phonetically-compact sentences (SI), designed as a compact set of sentences that collectively provide good diphone coverage; and phonetically diverse sentences (SX), selected to maximize the number of allophonic contexts in the text. Table 1 shows the distribution of sentence texts among speakers. In addition to the acoustic speech

Table 1: Distribution of TIMIT sentence texts

Sentence Type	Sentences	Speakers per Sentence	Total Utterances	Sentences per Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total:	2342	N/A	6300	10

Table 2: Distribution of TIMIT speakers by dataset.

Dataset	Male Speakers	Female Speakers	Total
TRAIN1	260	108	368
TRAIN2	66	28	94
TEST	112	56	168
Total:	438	192	630

data, each utterance in the TIMIT corpus contains a professionally produced, hand-labeled phonetic transcription that is aligned with the speech.

The speech utterances were recorded in a double-walled sound booth using a Sennheiser HMD-414 head-mounted, noise-cancelling microphone and stored at a 16 kHz sampling rate. The corpus is divided into TRAIN and TEST sets containing independent sets of speakers. As the present work requires an additional independent dataset for validation, the TRAIN set was subdivided into two subsets as follows: 80% of the speakers in each gender were randomly selected from the TRAIN set and assigned to the TRAIN1 set, while the remaining 20% of the speakers were assigned to the TRAIN2 set. Table 2 lists the number of male and female speakers in each dataset. Each dataset contains all 10 sentences for each of its speakers.

CHAPTER V

INVERSE FILTERING METHODS

A general discussion of the concept and limitations of inverse filtering was given in Section 3.1.3. The purpose of this chapter is to describe the four algorithms that were selected for this study, which represent a set of important ideas from the inverse filtering literature. Closed-phase linear prediction analysis (Section 5.1), one of the oldest automatic inverse filtering algorithms, relies on the concept of source-tract separability during the closed glottal phase, an idea that is still being exploited by newly proposed algorithms [10]. Glottal quality inverse filtering (Section 5.2) somewhat relaxes the assumptions about the closed phase and instead estimates the vocal tract filter from a short time region *in the vicinity* of the closed phase, chosen according to a set of quality measures on the candidate glottal waveforms. The iterative adaptive inverse filtering algorithm (Section 5.3) is a widely used method that relies on preprocessing the speech signal to remove the (approximate) spectral influence of the glottal waveform and then estimating the vocal tract filter using data from the entire speech frame. Finally, the recent algorithm by Fu and Murphy [44] (Section 5.4) was selected as an example of a recent, state-of-the-art algorithm that incorporates a parametric glottal model into the estimation process while rejecting the assumption of a time-invariant vocal tract.

5.1 Closed Phase Linear Prediction Analysis

Closed-phase inverse filtering (CPIF) [121] is a classical procedure for extracting glottal waveform estimates from the speech signal automatically and objectively. The guiding principle behind CPIF is that the closed glottal phase is the optimal time

segment of the glottal cycle for performing VTF estimation, as the air from the sub-glottal region is greatly reduced or cut off, implying reduced physical interaction between the vocal tract and sub-glottal regions and therefore maximizing the validity of the convolutional source-filter model given in Equation 1. In addition, glottal closure tends to be abrupt, so that the time-domain speech signal $s[n]$ should be most similar to the impulse response $v[n]$ of the vocal tract in a time interval that begins at the instant of glottal closure and extends through the closed glottal phase.

The challenge in obtaining a good implementation of CPIF lies in being able to automatically detect the closed glottal phase directly from the speech signal, which is rather difficult since an estimate of the glottal waveform has yet to be computed. The original implementation by Wong, Markel and Gray [121] was based on finding a minimum region of the normalized linear prediction error residual. The implementation used in this study estimates glottal closure instants using the recently developed DYPSA algorithm [70, 85] as implemented in the VOICEBOX toolbox [19].

For each frame of speech $s_k[n]$ with at least one GCI estimate, the CPIF algorithm proceeds as follows: The starting sample of the length $N_{lpa} = 2\,p + 1$ analysis window

$$\mathbf{w}_{ap} = \left[s_k[\hat{n}]\, s_k[\hat{n}+1] \ldots s_k[\hat{n}+N_{lpa}-1] \right]^T,$$

where p is the model order, is aligned with the GCI estimate $\hat{n}$ closest to the center of the frame. A least-squares, all-pole estimate of the vocal tract filter

$$\hat{V}_k(z) = \frac{1}{1 + \sum_{m=1}^{p} a(m) z^{-m}} \tag{14}$$

is then computed from $\mathbf{w}_{ap}$ using the covariance method of linear prediction analysis (LPA) [58]. The covariance method of linear prediction is not guaranteed to produce a stable filter. To address this issue, poles of $\hat{V}_k(z)$ lying outside the unit circle are reflected as a post-processing step to enforce filter stability. The reflection operation does not affect the magnitude spectrum of the filter. Similarly, because the vocal

tract is not supposed to have resonances at DC, positive real poles are removed from $\hat{V}_k(z)$.

Given a vocal tract filter estimate $\hat{V}_k(z)$, a first-order lip-radiation filter $R(z) = 1 - \alpha\, z^{-1}$, and assuming the linear source-filter model of speech production of Equation 1, the glottal waveform estimate $\hat{g}_k[n]$ is obtained as follows:

$$\hat{g}_k[n] = \left(s_k[n] + \sum_{m=1}^{p} a(m) s_k[n-m] \right) + \alpha\, \hat{g}_{ki}[n-1]. \tag{15}$$

5.2 Glottal Quality Inverse Filtering

The main idea behind the Glottal Quality Inverse Filtering (GQIF) algorithm is to extend closed-phase inverse filtering to situations where the estimation of the glottal closure instant (GCI) is inexact, or where the closed phase is so short that placing the spectral analysis window at the GCI would cause it to extend beyond the point of glottal opening. This is achieved by allowing the spectral analysis window to shift around the GCI estimate, producing several candidate glottal estimates. A ranking procedure is then employed to select the "best" estimate as the final solution.

Because the true glottal waveform is not directly observed in the acoustic speech signal, the *quality* of a glottal waveform estimate is a vague concept at best. Nevertheless, a set of criteria allowing for an approximate, objective quality assessment of a given set of glottal waveform estimates can enable the design of a more robust glottal waveform estimation algorithm. Several proposed glottal waveform quality measures (GQMs) [13, 5, 80] may be used for this purpose. A GQM is defined as a scalar-valued function of the glottal waveform estimate (and/or possibly, the associated vocal tract filter) that outputs a numerical value representing the relative quality of inverse filtering, such that values closer to either extremum of the real line indicate "better" or "worse" glottal waveform estimates, respectively. As it is unlikely that any single GQM can completely assess the quality of a glottal waveform estimate, Moore and Torres [80] presented a rank-based method to allow the combination of

an arbitrary number of GQMs for the selection of the "best" out of a set of glottal waveform estimates. Using sustained vowel recordings for which a simultaneous EGG signal was available, a later study by the same authors [81] evaluated the performance of all possible combinations of a set of 12 GQMs. Parting from the assumption that covariance-LPA is expected to perform better when the analysis window is near the region of glottal closure, the authors defined GQM performance as the frequency with which a particular GQM subset selected a mostly closed-phase glottal waveform estimate over a mostly open-phase estimate. The beginning of closed and open glottal phases were approximated a priori using EGG signals (Section 3.1.2). The study found that a combination of four GQMs could correctly select glottal waveform estimates obtained using analysis windows located in the closed glottal phase 94.7% of the time.

The following sections describe the GQIF algorithm, which is based on the results in [81]. The presentation of the rank-based method used to select an "optimal" glottal waveform estimate is followed by a brief discussion of the four GQMs used by the algorithm, and finally, a formal description of the GQIF procedure.

5.2.1 Rank-Based Glottal Waveform Quality Assessment

As mentioned in the previous section, choosing the "best" $g_k[n]$ estimate based on any single GQM is inherently dependent on the extremum of the measure (e.g., the maximum or minimum value), which can vary or become ambiguous if there is noise or quantization of the GQM values. It seems more natural to assume that good GQMs should reliably establish *trends* among a set of glottal waveform estimates (i.e., from relatively good to relatively bad) without the "best" or "worst" necessarily being represented by the extreme values. Additionally, no single GQM is designed to measure *all* of the qualities of a glottal waveform estimate and it is likely that a combination of GQMs would produce better results. In [80], Moore and Torres

introduced a simple new technique for combining multiple GQMs to evaluate the relative quality of a set of $g_k[n]$ estimates. This technique is referred to as Rank-Based Glottal Waveform Quality Assessment (RB-GQA) and it is implemented in the following steps:

1. For each GQM, rank each stored estimate from '1' to the number of stored estimates available (i.e., for N stored estimates, a rank of '1' indicates the "best" of the stored estimates for that GQM and rank of N indicates the worst).

2. Compute the average ranking across a subset of GQMs and sort the estimates by increasing average rank. The estimate with the lowest average rank (i.e., closest to 1) is selected as the highest-quality estimate.

An advantage of RB-GQA is that it is invariant to any monotonic transformation of the GQM values, thus allowing input from any subset of the GQMs to be effectively combined in the final quality assessment.

5.2.2 Glottal Quality Measures

5.2.2.1 Phase-Plane Measures

The work in [13] presented two GQMs based on phase-plane analysis. These measures rely on the assumption that the glottal waveform can be modeled as a second-order harmonic equation, which implies that its plot in the phase-plane $(x(t), \frac{dx}{dt})$ should consist of one closed loop per fundamental period. Resonances not completely removed by inverse filtering should appear as sub-cycles within the fundamental loops. The phase-plane plots are quantified by measures reflecting the number of cycles per fundamental period (pp_{cper}), with fewer cycles reflecting better estimates, and the mean sub-cycle length (pp_{cyc}), with smaller sub-cycles reflecting better estimates. These measures are computed using MATLAB code from TKK Aparat [1]. Figure 5 shows the phase-plane plots of a higher and lower-quality glottal waveform estimate, respectively.

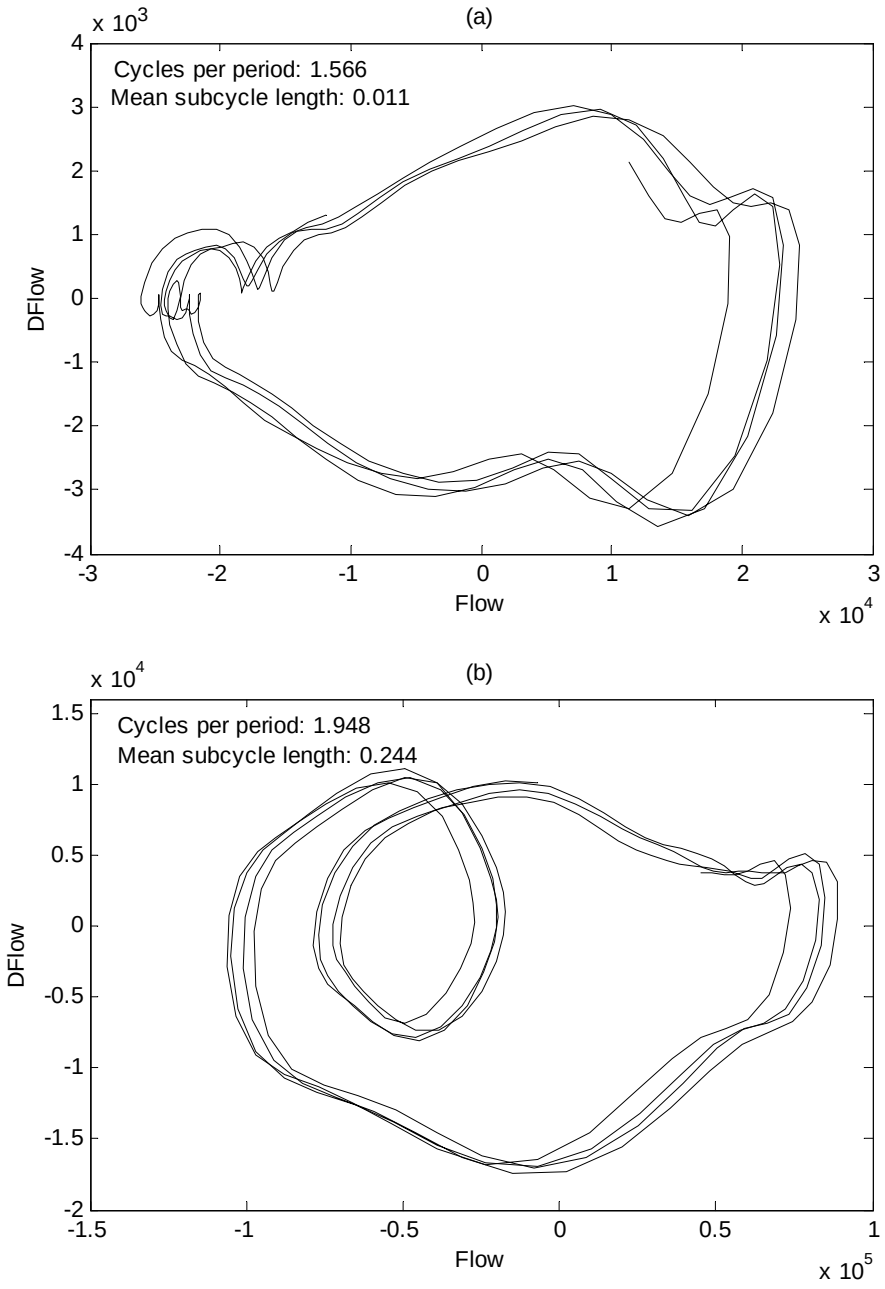


Figure 5: Phase-plane plots and corresponding GQMs for (a) the higher quality estimate in Figure 4(a), and (b) the lower quality estimate in Figure 4(b). The lower quality estimate shows an additional large sub-cycle.

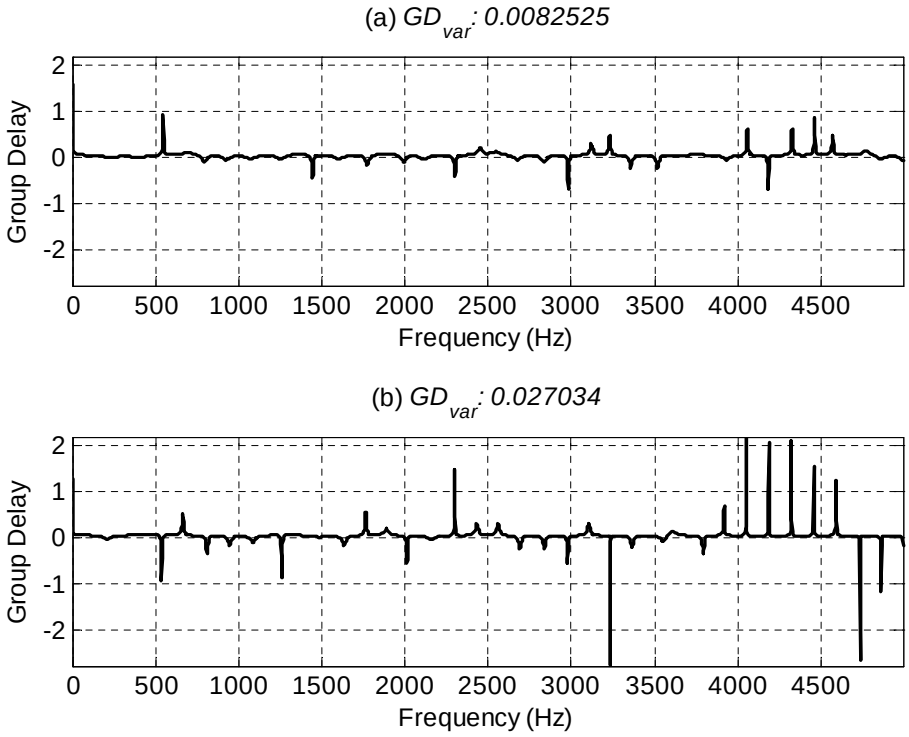


Figure 6: Group delay and its variance for (a) the higher quality estimate in Figure 4(a), and (b) the lower quality estimate in Figure 4(b). The group delay of the lower quality estimate shows additional extraneous peaks that increase the variance.

5.2.2.2 Group Delay

The motivation for using the group delay as a GQM was presented in [5], where it was observed that the phase spectrum over a single cycle of the glottal flow should be essentially constant over a wide frequency range if the vocal tract resonances were completely removed by the inverse filtering procedure. RB-GQA uses the variance of the group delay (GD_{var}) of the glottal flow (computed over a single cycle and using an FFT size of 4096) as a GQM. Better estimates of the glottal waveform are be expected to have a variance closer to zero. Figure 6 shows the group delay function for a higher and lower-quality glottal waveform estimate and their corresponding variances.

5.2.2.3 Harmonic Ratio

Ideally, the spectrum of the glottal waveform $G_k(e^{j\omega})$ should exhibit a strictly negative spectral slope due to the lack of resonant structure. If formant residuals from an improperly estimated $\hat{V}(z)$ are present, this linear trend is disturbed. RB-GQA uses a GQM based on the ratio of the first harmonic peak to the maximum peak present over a frequency range 0 – 3700 Hz (hr_{mx}). Ideally, the first harmonic peak *should* tend to be greater than successive peaks to adhere to the negative linear trend expected from an ideal glottal waveform. Deviations from this trend can create ratios that are greater than one and indicate worse glottal waveform estimates. The frequency range of 0–3700 Hz was used to cover the most prominent formants in voiced speech. Figure 7 shows examples and the resulting measurements of the harmonic ratio and linear regression GQMs for a lower and higher-quality glottal waveform estimate.

5.2.3 GQIF Algorithm

Given a frame of speech $s_k[n]$ and a GCI estimate $n_{e,k}$ (In current implementation, the GCIs are estimated from $s[n]$ using the DYPSA algorithm), the GQIF Algorithm proceeds as follows:

1. Let $N_{lpa} = 2p + 1$ be the length of the LP analysis window, where p is the LPA model order.

2. For $i = n_{e,k} - N_{lpa} + 1 \ ... \ n_{e,k}$

   (a) Compute the all-pole vocal tract estimate

   $$\hat{V}_i(z) = \frac{1}{1 + \sum_{m=1}^{p} a(m) z^{-m}},$$

   via covariance LPA, using analysis window

   $$\mathbf{w}_i = [s_k[i] \, s_k[i+1] \, ... \, s_k[i + N_{lpa} - 1]]^T .$$

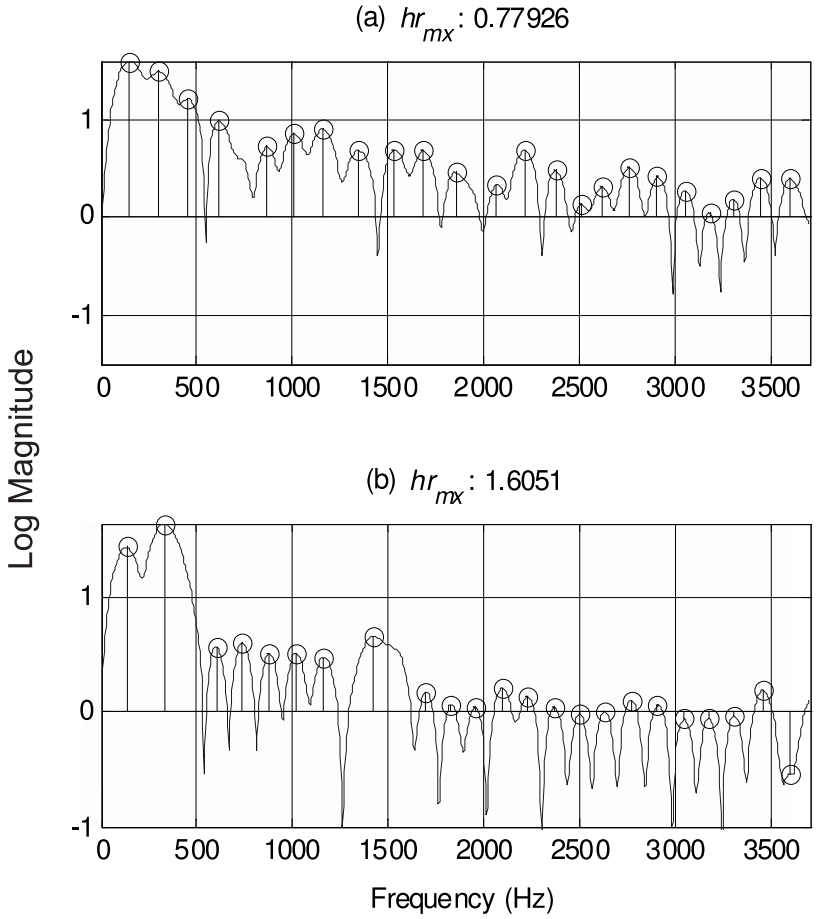


Figure 7: Harmonic peaks and corresponding hr_{mx} values for (a) the higher quality estimate in Figure 4(a), and (b) the lower quality estimate in Figure 4(b). The lower quality estimate has a second harmonic that is larger than the first harmonic, and a large peak around 1500 Hz, causing an increase in the hr_{mx} measure.

(b) Compute the glottal waveform estimate via inverse filtering:

$$\hat{g}_{ki}[n] = \left(s_k[n] + \sum_{m=1}^{p} a(m) s_k[n-m] \right) + \alpha\, \hat{g}_{ki}[n-1],$$

where α is the lip radiation coefficient.

(c) Calculate and store the value of each GQM for $\hat{g}_{ki}[n]$:

$$pp_{cper}[i], pp_{cyc}[i], GD_{var}[i], hr_{mx}[i].$$

3. Use the GQMs and the RB-GQA procedure described in Section 5.2.1 to find the optimal value of the analysis window position $\tilde{i}$. Return the final estimate $g_k[n] = g_{k\tilde{i}}[n]$.

5.3 Iterative Adaptive Inverse Filtering

An alternative to finding an optimal short region of the pitch cycle from which to perform spectral estimation of the vocal tract is to estimate the vocal tract filter from an entire frame of speech that has been pre-processed to approximately remove the spectral influence of the glottal source. The iterative adaptive inverse filtering (IAIF) algorithm is based on this idea. In this study, the IAIF algorithm was used as implemented in the TKK Aparat toolbox [1]. Although the algorithm is fully specified in this section, the exact implementation may be obtained from the Aparat website [2]. The algorithm is described in detail in the original paper by Alku [8] and in the Aparat documentation [1].

IAIF is based on a spectral estimation method called discrete all-pole modeling (DAP) [38], which was designed to overcome the biasing of VTF pole estimates towards pitch harmonics that occurs in autocorrelation LP analysis of high-pitched speech. The IAIF procedure is a multi-stage process in which progressively detailed estimates of the glottal waveform's spectrum are removed from the speech spectrum to obtain progressively refined estimates of the vocal tract filter. It should be noted that the original paper by [8] recommends high-pass filtering the speech signal with

a high-pass filter having a cutoff frequency well below f_0 as a pre-processing step to remove low-frequency fluctuations in the signal. That step is not implemented by the `iaif` function in the current version of Aparat [2], nor was it deemed necessary for this study since the analyzed speech data was captured via a high-quality recording process (Section 4.6) and there are no low-frequency noise/bias issues associated with it. The IAIF algorithm for a frame of speech $s_k[n]$ is described below. For notational simplicity, the description is given in the Z domain, although the filtering operations in IAIF are actually performed in the time domain. A fixed, first-order lip radiation filter $R(z) = 1 - \alpha\, z^{-1}$ is used throughout.

1. Fit a 1^{st}-order all-pole filter $H_{g1}(z)$ to $S_k(z)$ via DAP. $H_{g1}(z)$ represents a first approximation of the glottal waveform's spectral tilt.

2. Compute $S_{g1}(z) = S_k(z)/H_{g1}(z)$, a processed version of the speech frame from which the glottal spectral tilt has been approximately removed.

3. Fit a p^{th}-order all-pole filter $V_1(z)$ to $S_{g1}(z)$ via DAP. This is the first approximation of the vocal tract filter.

4. Compute $G_1(z) = S_k(z)/(V_1(z)R(z))$, the first estimate of the glottal waveform.

5. Fit an 4^{th} order all-pole model $H_{g2}(z)$ to $G_1(z)$ using DAP. $H_{g2}(z)$ is a refined parametric model of the glottal waveform's spectrum.

6. Compute $S_{g2}(z) = S_k(z)/(H_{g2}(z)R(z))$, a processed version of the speech frame from which the spectral effects of the glottal waveform have been mostly removed.

7. Fit a p^{th}-order all-pole filter

$$\hat{V}(z) = \frac{1}{1 + \sum_{m=1}^{p} a(m) z^{-m}}$$

to $S_{g2}(z)$ via DAP. This is the final approximation of the vocal tract filter.

8. Compute the final estimate of the glottal waveform as

$$\hat{G}_k(z) = S_k(z)/(\hat{V}(z)R(z)).$$

The time-domain expression describing this final inverse filtering step is identical to Equation 15.

5.4 Time-varying Inverse Filtering via the Fu-Murphy Algorithm

Recent work on inverse filtering algorithms has focused on relaxing some of the assumptions about the speech signal made by previous procedures in an effort to produce better estimates of the glottal waveform. A representative of this research direction is the algorithm by Fu and Murphy [44], which relaxes the assumption of a stationary vocal tract by incorporating a time-varying vocal tract filter model into the inverse filtering process. The use of the Fu-Murphy inverse filtering algorithm (FMIF) in this study is based on an implementation of the algorithm described in [44] by the author of this thesis, since it was not possible to obtain a copy of the original implementation. This section provides provides a functional description of the algorithm implementation used in this study and explains any differences to the description in the original paper when they arise.

The FMIF algorithm works pitch-synchronously and is based on a time-varying model of the vocal tract. For a speech frame $s_k[n]$, the instants of glottal closure (GCIs) (estimated by the DYPSA algorithm in this author's implementation) are used to segment each pitch period. For each pitch period $s_r[n]$, the algorithm begins by finding an approximate estimate of the VTF and glottal waveform. While the original paper by Fu and Murphy describe the use of pitch-synchronous LP-analysis on pre-emphasized speech to obtain a preliminary (time-invariant) VTF estimate and the least-squares fitting of a Rosenberg model to obtain an initial glottal waveform estimate, the implementation used in this study simply uses the VTF $\hat{V}_k(z)$ and

glottal waveform estimate $\hat{g}[n]$ obtained via IAIF for the entire frame as initial values for the FMIF optimization.

The FMIF algorithm consists of a descent optimization procedure whereby the residual of a time-varying autoregressive vocal tract filter that is estimated from the speech signal using a Kalman filter is compared to a potential parametric estimate of the glottal waveform. Given a segment of speech $s_r[n]$ representing a single pitch period, the FMIF algorithm seeks to represent the $s_r[n]$ as follows:

$$s_r[n] = g'[n] - \sum_{i=1}^{p} a_i[n] s_r[n-i], \tag{16}$$

where $a_i[n]$ are the time-varying coefficients of the autoregressive (all-pole) vocal tract filter and $g'[n]$ is the glottal waveform derivative, which is represented by the Liljencrants-Fant (LF) model [41]. The objective function to be minimized is given by

$$\mathcal{E}(\theta) = \left\| s_r[n] - g'_r[n] + \sum_{i=1}^{p} a_i[n] s_r[n-i] \right\|^2, \tag{17}$$

where θ represents the LF-model parameter vector that uniquely determines $g'_r[n]$. An optimal value of θ that minimizes $\mathcal{E}(\theta)$ is found using the interior-trust-region descent algorithm described in [26] and implemented by the `fmincon` function in MATLAB. On every evaluation of the error function $\mathcal{E}(\theta)$, the time-varying VTF coefficients are obtained by the Kalman filtering procedure described in [44]. The final result is a set of LF-model parameters for synthesizing $g'_r[n]$.

CHAPTER VI

ACOUSTIC SPEECH FEATURES AND FEATURE EXTRACTION

6.1 Spectral Envelope Features

A widespread approach for obtaining a useful set of features for speech analysis is via feature extraction procedures inspired by human auditory perception. It has been found that the goal of obtaining a compact representation of the speech signal that retains information which is relevant to the analysis task at hand can be met, at least in the case of speech recognition and speaker identification, by a spectral magnitude envelope representation in a warped frequency domain that is consistent with experimentally derived frequency resolution properties of the human auditory system [49]. This section describes the spectral envelope features (SEFs) used in this study: the quasi-ubiquitous [126, 94] mel-frequency cepstral coefficients ($mfcc$); Perceptual linear prediction (plp), which includes additional processing steps to model the amplitude dynamics of auditory perception, and a decorrelated mel-scale subband feature set ($melsub$) that has been found to produce very high classification rates in speaker identification [77] and speech recognition [84, 90].

The extraction of spectral envelope features (SEFs) typically begins with the construction of a perceptually-spaced bank of band-pass filters. A common choice for the spacing of these filters is based on the mel-scale, defined in [125] as

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}). \tag{18}$$

The mel-scale is loosely based on perceptual experiments which have found that the frequency-resolution of human hearing is higher at lower frequencies and decreases logarithmically at higher frequencies. A mel-scale filter bank is constructed by equally

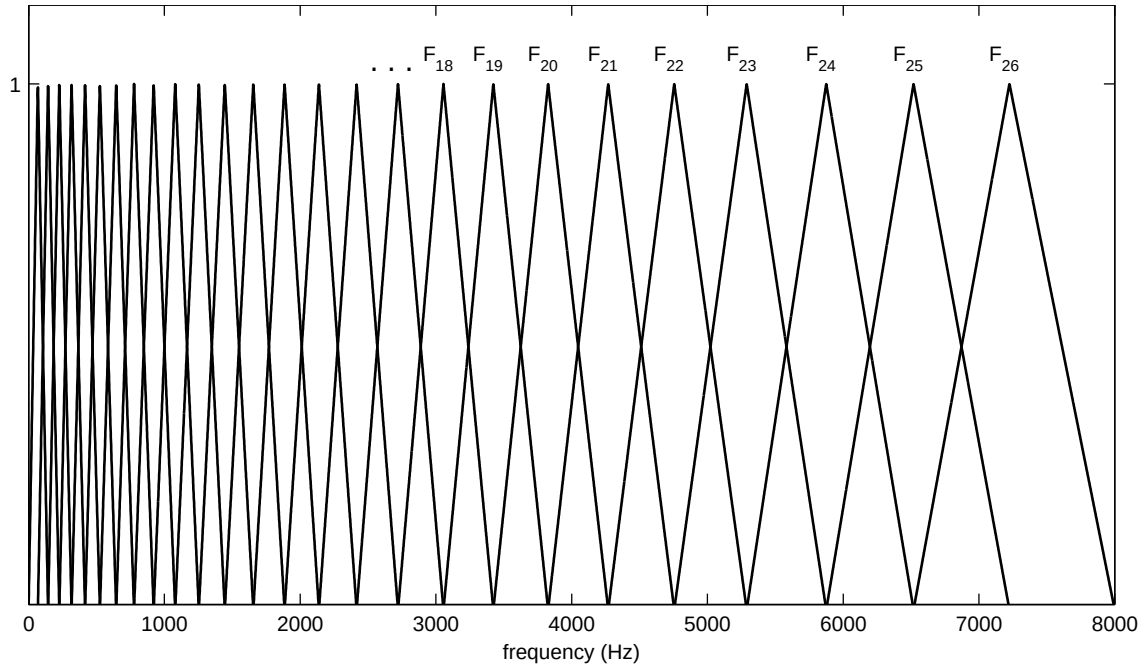


Figure 8: 26-channel mel-scale filter bank.

spaced B triangular filters along the mel-frequency axis. The bandwidth of the b^{th} filter $F_b(\omega)$ is such that it extends from the center frequency of $F_{b-1}(\omega)$ to the center frequency of $F_{b+1}(\omega)$, as shown in Figure 8. Each filter is applied by weighting each frequency bin in the input spectrum by the magnitude of F_b at that frequency. The output of the filter bank is a signal $\Psi[b]$ representing the spectral energy at each filter bank channel b. For a given M-point discrete spectrum $S_k[m]$ (i.e. the DFT of speech frame $s_k[n]$ after modulation by a Hamming window), Ψ is computed as follows:

$$\Psi[b] = \sum_{m=0}^{M-1} \left| S_k[m]\, F_b(2\pi m/M) \right|^2 . \tag{19}$$

The filter bank energies $\Psi[b]$ can be considered to be a low-dimensionality representation of the power spectrum designed to hide fine spectral structure while highlighting broader spectral variations that way be perceptually relevant. The final SEF vector is obtained through additional processing of $\Psi[b]$ to further reduce dimensionality and/or decorrelate the B channel outputs. In this study, spectral envelope features were obtained using the `HCopy` program included with HTK version 3.4 [125].

6.1.1 Mel-Frequency Cepstral Coefficients

Mel-frequency cepstral coefficients were originally devised by Davis and Mermelstein [30] with the objectives of obtaining a compact representation of the speech signal that suppressed non-phonetic information while enhancing aspects of the signal pertinent to speech recognition. The mel-frequency cepstrum is computed by applying a discrete cosine transform (DCT) to the log mel-scale filter bank energies $\Psi[b]$. It is defined in [125] as

$$C[l] = \sqrt{\frac{2}{B}} \sum_{b=1}^{B} \log\left(\Psi[b]\right) \cos\left(\frac{\pi\, l}{B}\left(b - 0.5\right)\right), \tag{20}$$

where l denotes cepstral coefficient index. The purpose of transforming the filter bank energies to cepstral coefficients is two-fold: First, the DCT, which is close to the Karhunen-Loève transform, results in approximately decorrelated features. Second, a property of the cepstrum is that the broad spectral variations are encoded by the low-order coefficients, while the spectral details are represented by the higher-order coefficients. Thus, by selecting the first L coefficients, where $L < B$, a smoother representation of the spectral envelope can be obtained. For the purposes of this study, values of $L = 13$ and $B = 26$ were used to compute $mfcc$ SEFs. These are fairly typical settings for a 16 kHz sampling rate. The resulting L-dimensional $mfcc$ vector includes the 0^{th} cepstral coefficient.

6.1.2 Perceptual Linear Prediction

Perceptual linear prediction was introduced by Hermansky [61] as a compact representation of the speech spectrum that modeled the following psychoacoustic properties: variable frequency resolution, variable loudness perception as a function of frequency, and the power law relationship between intensity and loudness. While the original implementation is based on the (similar) Bark scale, the HTK implementation of plp feature extraction is based the same mel-frequency filter bank used for $mfcc$. The mel-scale filter bank energies $\Psi[b]$ are weighted by an equal-loudness curve and

then compressed by taking the cubic root as a way to model the amplitude dynamics of human hearing. From these modified filter-bank energies $\hat{\Psi}[b]$, further frequency smoothing and dimensionality reduction is achieved by computing the coefficients $A[q]$ of a Q^{th}-order all-pole model using the autocorrelation method of linear prediction (LP). To obtain approximately decorrelated features, the LP coefficients are transformed to cepstral coefficients by the recursion

$$C[q] = -A[q] - \frac{1}{q}\sum_{i=1}^{q-1}(q-i)A[i]\,C[q-i]. \tag{21}$$

The LP order was set to $Q = 12$, and a 26-channel filter bank was used as with the $mfcc$ features.

6.1.3 Decorrelated Mel-Frequency Filter-Bank Energies

Although cepstral features such as $mfcc$ and plp have a long usage history in speech analysis, studies by Nadeu et al. [84] and Paliwal [90] have independently found that if some simple post-processing is performed on the log filter-bank energies $\log(\Psi[b])$, these features can meet or even improve upon the performance of $mfcc$ features on speech recognition. These results have been recently validated in [77], where features obtained following the approach of [84] were found to outperform $mfcc$ features in speaker identification. For the purposes of this study, the decorrelated mel-frequency filter bank energy features ($melsub$) are computed following [77], where $B = 21$ filter-bank channels were used to process 16 kHz speech. The $melsub$ features

$$D[b] = \log(\Psi[b]) - \log(\Psi[b-1]) \tag{22}$$

are obtained from $\Psi[b]$ by taking a first-order difference across the channel index b. This simple procedure was found in [84] to approximately decorrelate the resulting feature set. A practical advantage of this feature set over cepstral coefficients is that the $D[b]$ coefficients have a clear physical interpretation. Each coefficient $D[b]$ can be regarded as a local spectral slope, whose bandwidth varies in a psychoacoustically

correct way across frequency. The frequency-domain locality of the coefficients has been exploited in [77] in a speaker ID application on noisy speech, where, in a given speech frame, the noise may corrupt only some of the frequency channels. In this study, a higher-resolution version of this feature set using a 41-channel filter bank ($melsub_{41}$) is also investigated.

6.2 Glottal Waveform Features

Various measures taken from the glottal waveform estimates obtained via inverse filtering have been proposed as useful ways for objectively characterizing the pattern of airflow across the glottis. Such measures may be obtained by one of three approaches: (1) direct measurement of the time-domain glottal waveform, (2) direct measurement of the waveform's magnitude spectrum, or (3) fitting a parametric model. The exact procedure for performing each measurement can vary by author, and there is no strong consensus on the exact definition of some measurements. Nevertheless, glottal waveform measures can be generally grouped according to the underlying aspect of glottal behavior that they intend to quantify. This section describes existing glottal feature categories and defines the feature set used in this study.

6.2.1 Salient Features of the Glottal Cycle

A primary feature of the glottal cycle is the duration of the time interval where the vocal folds are open and air flows through the glottis, denoted as the *open phase*, in relation to the duration of the *closed phase*, where airflow is generally cut off. The "duty cycle" of the glottal waveform is measured by the **open quotient (OQ)**, which is defined as the ratio between the duration of the open phase to the duration of the glottal cycle. A second salient time-domain feature is the asymmetry of the open phase, which is usually quantified by the **speed quotient (SQ)**, a ratio between the duration of the *opening phase* and *closing phase* of the vocal folds. For non-pathological, modal phonation, the vocal folds open more slowly than they close,

resulting in an SQ value greater than one. The **closing quotient** (ClQ), defined as the duration of the closing phase relative to the pitch period, may be used as an alternative time-domain measure since it does not depend on estimating the instant of glottal opening, which can be difficult to find when the glottal waveform estimate is corrupted by formant ripple due to errors in inverse filtering. The **return quotient** (Qa) measures the effective duration of the *return phase*, which starts at the point of maximum *closing* (i.e. the glottal closure instant) and continues up to the point of maximum glottal *closure*, where airflow through the glottis is minimal or non-existent.

One of the most perceptually salient glottal effects on the speech signal is the introduction of a spectral roll-off in the frequency domain that is approximately linear on a log-log scale over much of the frequency range. This effect is measured from the glottal waveform spectrum by the **spectral tilt** ($TILT$), which is defined as the slope of spectral roll-off starting from the spectral peak (glottal formant), generally located between f_0 and $2f_0$ (i.e., between the first two harmonics). The **harmonic richness factor** (HRF) [21], defined by the spectral amplitude of the higher harmonics relative to the amplitude of the first harmonic, is a closely related measure which was proposed for its ability to characterize variations in the high-frequency portion of the spectrum due to different phonation types.

Research efforts on glottal waveform parameterization have also sought to quantify glottal behavior in a meaningful way by a single parameter. The **harmonic level difference** ($H1$-$H2$) has been widely used to characterize phonation type and voice quality [69, 55, 56, 108]. In [40], a regression analysis of vowels along a pressed-breathy phonation continuum led to the proposal of the **shape parameter** (Rd), which is closely related to the rate of glottal closure and was found to be effective as a single-parameter descriptor of pressed or breathy voice quality on a set of Swedish vowels. Interestingly, this parameter was found to have a near-perfect linear relationship to $H1$-$H2$, and was later discovered in [35] to be closely related to a perceptual distance

measure.

In the context of the present study, it is important to understand that the aforementioned time-domain glottal measures have theoretically known frequency domain correlates [40, 108, 59, 12, 36]. The analysis of synthetic glottal waveforms generated by the Liljencrants-Fant (LF) model reveals that Qa controls the starting frequency of an additional -6 dB/octave of spectral roll-off. This starting frequency is lower for a longer return phase. Meanwhile, open and speed quotients were found to modify the center frequency and bandwidth, respectively, of the spectral peak, which is located at a lower center frequency for higher OQ and has a larger bandwidth for higher SQ. Therefore, some amount of correlation is expected to exist between OQ (or SQ) and $H1$-$H2$, as well as Qa with $TILT$ and HRF. However, each of these related features will also reflect the advantages and disadvantages of the measurement method (time-domain, frequency-domain, or model fitting) through which they are obtained.

6.2.2 Measurement Methods

In what follows, the 16 glottal features used in this study are formally defined. This feature set was chosen to include, where applicable, several variations of each feature type, obtained via different measurement methods. The reader is referred to Figures 9 and 10 for illustration of the time instants, amplitude intervals, and pitch harmonics described in the text.

6.2.2.1 Time-Domain Direct Waveform Measurement

Direct measurement time-domain features were computed using modified code from the TKK Aparat toolbox [1]. The `glottaltimeparams` function was modified to use glottal closure instant (GCI) estimates to segment individual pitch cycles from frame-based glottal waveform estimates and to compute additional amplitude-threshold based variations of OQ and SQ. Traditionally, the estimation of time-domain features is based on the detection of critical time instants that are intended to denote

the boundaries between distinct phases of the glottal cycle. To reduce quantization errors in the detection of these time instants, which arise due to a finite sampling rate, cubic-spline interpolation was used to refine the detected time instants to fractions of a sample. Estimates of the glottal closure instants t_e had been previously obtained using the DYPSA algorithm [70, 85] as implemented in the VOICEBOX toolbox [19].

For each glottal cycle k, the fundamental frequency $f_{0,k} = 1/T_{0,k}$, where $T_{0,k} = t_{e,k} - t_{e,k-1}$, was computed from the time difference between two adjacent GCI's, and the instant of maximum *closure* $t_{c,k}$, which denotes the start of the closed phase, was detected as the first zero crossing of the glottal waveform derivative $g'[n]$ after $t_{e,k}$. The time instant of maximum airflow $t_{p,k}$, which denotes the boundary between the opening and closing phases, was computed as the local maximum of the glottal waveform $g[n]$ in the interval between $t_{e,k-1}$ and $t_{e,k}$.

The point of minimum airflow $t_{min,k}$ was computed by finding the local minimum of $g[n]$ in the interval between $t_{c,k-1}$ and $t_{p,k}$, and the instant of maximum *opening* $t_{\Delta max,k}$ was found as the location of the maximum value of $g'[n]$ between $t_{min,k}$ and $t_{p,k}$. The amplitude of glottal airflow was then computed $A_{ac,k} = g[t_{p,k}] - g[t_{min,k}]$.

To detect the instant of glottal opening, the glottal waveform derivative was first scanned starting from $t_{min,k}$ to find the point where it crossed a threshold of 10% above $g[t_{min,k}]$ relative to $A_{ac,k}$, denoted as $t_{10\%,k}$. The primary opening instant $t_{o1,k}$ was then detected as the *last* positive zero crossing of $g'[n]$ between $t_{min,k}$ and $t_{10\%,k}$. Due to ambiguities in the detection of the glottal opening instant in (noisy) inverse-filtered glottal waveforms, a secondary opening instant $t_{o2,k}$, proposed in [92], was computed as the local maximum of a smoothed version of the second glottal waveform derivative $g''[n]$ in the interval between $t_{o1,k}$ and $t_{p,k}$.

From these time instants, Aparat computes the durations of the *primary opening phase* $T_{op1,k} = t_{p,k} - t_{o1,k}$, *secondary opening phase* $T_{op2,k} = t_{p,k} - t_{o1,k}$, and *closing phase* $T_{cl,k} = t_{c,k} - t_{p,k}$, which leads to the calculation of the primary and secondary

speed quotients (SQ_1, SQ_2) and the closing quotient (ClQ), defined as follows:

$$SQ_1 = \frac{T_{op1}}{T_{cl}} \tag{23}$$

$$SQ_2 = \frac{T_{op2}}{T_{cl}} \tag{24}$$

$$ClQ = \frac{T_{cl}}{T_0}. \tag{25}$$

Because exact estimation of critical time instants via peak detection can be problematic when performed on noisy IF-derived glottal waveform estimates, measures have been proposed based upon the measurement of time intervals between amplitude threshold crossings. The threshold-based opening and closing instants were computed from the 20%, 50%, and 80% positive and negative threshold crossings of $g[n]$, respectively, as shown in Figure 9. From these instants, variations of OQ and SQ were computed as follows:

$$SQ_{20\text{-}80} = \frac{t_{o80} - t_{o20}}{t_{c20} - t_{c80}} \tag{26}$$

$$OQ_{20} = \frac{t_{c20} - t_{o20}}{T_0} \tag{27}$$

$$OQ_{50} = \frac{t_{c50} - t_{o50}}{T_0} \tag{28}$$

$$OQ_{80} = \frac{t_{c80} - t_{o80}}{T_0}. \tag{29}$$

Another alternative to the exact detection of critical time instants has been to construct features based on amplitude level ratios. In an effort to produce more robust glottal features, the **normalized amplitude quotient** (NAQ) was introduced in

[9] as a noise-resilient alternative to the closing quotient. Similarly, the amplitude-based measure OQa was proposed in [47] as a robust alternative to the timing- or threshold-based open quotient. These measures were computed in Aparat as follows:

$$NAQ = \frac{A_{ac}}{T_0 \, A_{\Delta min}} \tag{30}$$

back

$$OQa = A_{ac} \left(\frac{\pi}{2A_{\Delta max}} + \frac{1}{A_{\Delta min}} \right) f_0 \tag{31}$$

6.2.2.2 Liljencrants-Fant Model Fitting

Time-domain glottal features may also be obtained by least-squares fitting of a parametric model to a glottal waveform estimate. The simultaneous advantage and drawback of computing features by this method is that the fitted model will represent a theoretically correct glottal cycle. Thus, model fitting may be helpful in obtaining robust parameter estimates in the presence on noise in the speech signal and/or distortion in the glottal waveforms, but this may also result in less "useful" features due to the fact that the parametric model is based on the properties of an ideal, simplified glottal flow that may not fully represent the glottal variations present in real speech signals.

The Liljencrats-Fant (LF) model [41] has been widely used to parameterize the glottal waveform, and has been shown to be functionally equivalent to (and in some cases a superset of) other popular parametric glottal flow models [36]. The LF model represents the glottal waveform derivative $g'_{lf}(t)$ as follows:

$$g'_{lf}(t) = \begin{cases} A_0 \, e^{\alpha t} \sin(\omega_g t) & 0 \leq t < t_e \\ \frac{-A_0}{\varepsilon t_a} \left[e^{\varepsilon (t - t_e)} - e^{-\varepsilon (t_c - t_e)} \right] & t_e \leq t \leq t_c, \end{cases} \tag{32}$$

where $0 < t < t_p$ and $tp < t < te$ represent the opening and closing phases, respectively, and $te < t < tc$ represents the return phase, as shown in Figure 2. The

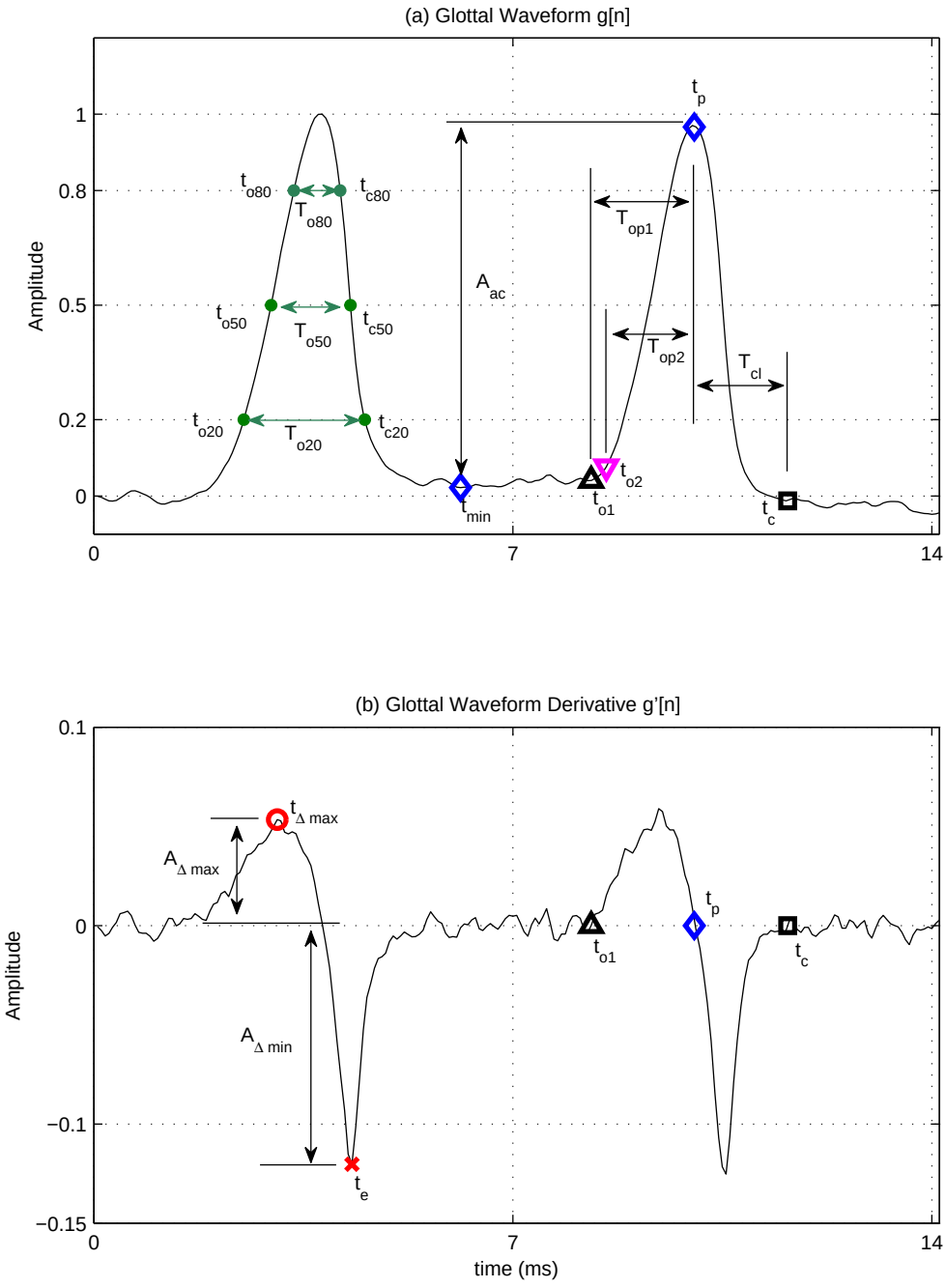


Figure 9: Annotated glottal waveform estimate (a) and its derivative (b), showing time instants and amplitude thresholds used for calculating time-domain glottal features.

synthesis parameters A_0, α, w_g, and ε are fully determined by the timing parameters t_p, t_e, t_a, t_c and $A_{\Delta min}$, as discussed in [41]. The parameter t_a is the projection of $g''_{lf}(t_e)$ onto the time axis, and controls the effective length of the return phase. For simplicity t_c can be set to T_0, and the model-fitting procedure reduces to finding the values of the four parameters t_p, t_e, t_a, and $A_{\Delta min}$ that minimize the mean squared error

$$\varepsilon_t = \sum_{n=0}^{N_0-1} \left| g'[n] - g'_{lf}(nT_s) \right|^2, \tag{33}$$

where $T_s = 1/f_s$ is the sampling period of the speech signal and $N_0 = \lfloor T_0 \, f_s \rfloor$ is the length of the glottal cycle in samples.

A solution to this non-linear minimization problem was found via an interior trust-region-reflective algorithm [26] as implemented by the `lsqnonlin` function in the MATLAB Optimization Toolbox. The constraints of the minimization problem can be simplified by transformation into a set of parameters with fixed bounds [36]. To this end, the timing parameters $\{t_p, \, t_e, \, t_a\}$ were transformed into the following set of equivalent parameters, which also form part of the glottal feature set:

$$OQ_{LF} = \frac{t_e}{T_0} \tag{34}$$

$$\alpha_m = \frac{t_p}{t_e} \tag{35}$$

$$Qa = \frac{t_a}{T_0 - t_e}, \tag{36}$$

where OQ_{LF}, defined between 0 and 1, is the LF-derived open quotient. The **asymmetry coefficient** (α_m), defined between 0.65 and 1, is related to the LF-model speed quotient by $\alpha_m = SQ_{LF}/(1 + SQ_{LF})$. The return quotient Qa ranges from 0 to 1 and controls the effective length of the return phase.

LF-model fitting was then performed by finding values of $\{OQ_{LF}, \, \alpha_m, \, Qa \;\; A_{\Delta min}\}$ that minimize ε_t, using the trust-region-reflective procedure. Initial values for the optimization algorithm were given by the direct-measurement estimates of t_p, t_e, and

$A_{\Delta min}$ described in Section 6.2.2.1. The initial value of Qa was set to 0.1. Before LF-fitting, $g'[n]$ was up-sampled by a factor of four to allow for accurate estimation of the parameters. Once the least-squares fit $\hat{g}'_{lf}(nT_s)$ is found, the maximum airflow A_{ac} was measured as shown in Figure 9 from the LF waveform. The shape parameter Rd was computed as follows:

$$Rd = \frac{A_{ac} f_0}{110 A_{\Delta min}}, \tag{37}$$

An important concern with the extraction of glottal features from glottal waveforms that have been estimated by inverse filtering is that the waveforms are subject to corruption due to phase distortion during the recording or transmission processes [120]. While this is not expected to be a major issue with the database used for this study, which was recorded with a high-quality microphone into digital media (Section 4.6), it is nevertheless worthwhile to explore the use of estimation methods that seek to circumvent the phase distortion problem. The LF-parameters may also be derived by matching the LF waveform's discrete magnitude spectrum to the discrete magnitude spectrum of the glottal waveform estimate [40]. This was implemented by replacing the objective function ε_t of Equation 33 with its magnitude spectrum counterpart:

$$\varepsilon_f = \sum_{m=0}^{M_{max}-1} \left(|G_d[m]| - |G_{dlf}[m]| \right)^2, \tag{38}$$

where G_d and G_{dlf} are the length N_0 DFTs of $g'[n]w[n]$ and $g'_{lf}[n]w[n] = g'_{lf}(n\,T_s)w[n]$, respectively, and $w[n]$ is an N_0 point Hanning window. The frequency range of the optimization was restricted to 0 – 4 kHz by setting $M_{max} = \lfloor 4000\, N_0\, T_s \rfloor$. The frequency-domain LF-fitting procedure was exactly the same as for the time domain, except that ε_t was replaced by ε_f.

6.2.2.3 Direct Spectral Magnitude Measurement

Glottal features resilient to phase distortion may also be obtained directly from the magnitude spectrum of $g'_{lf}[n]$. Aparat uses the magnitude spectrum of the entire

glottal waveform frame, as shown in Figure 10, and the magnitudes $\{H1, H2 \ldots H_N\}$ of the pitch harmonics (in dB) to compute the harmonic level difference ($H1$-$H2$) and the harmonic richness factor (HRF) as follows:

$$H1\text{-}H2 = H1 - H2 \tag{39}$$

$$HRF = \frac{\sum_{i>1} H_i}{H1}. \tag{40}$$

It should be noted that, to prevent small errors in f_0 estimation and natural cycle-to-cycle f_0 deviations from affecting the computation of these features, the harmonics are not measured exactly at integer multiples of f_0. Instead, the i^{th} harmonic is defined as the local maximum of the magnitude spectrum in the region $f_0\, i \pm f_0/2$ [1]. An additional frequency-domain measure, the spectral tilt (TILT) was computed from the magnitude spectrum on a log-log scale as the slope of the least-square line-fit in the frequency region $f_{max} - fs/2$, where f_{max} is the location of the global maximum of the magnitude spectrum and fs is the sampling rate. The TILT feature is given in units of dB/decade.

6.3 Feature Extraction

As explained in Section 3.1.3, existing inverse filtering algorithms rely on an all-pole representation of the vocal tract to estimate and remove its influence from the speech signal in order to reveal an estimate of the glottal waveform. Furthermore, three of the four inverse filtering algorithms being evaluated (Chapter 5) assume a stationary vocal tract across the extent of a speech frame. The assumption of an all-pole, locally time-invariant vocal tract is most optimally satisfied during the phonation of stationary vowels, as the production models for the other phonetic classes may introduce spectral zeros into the vocal tract model (nasals), add a noise source that is modulated by the glottal waveform (voiced fricatives), involve a shorter duration than that of a typical speech frame (plosives), or be characterized by a time-varying

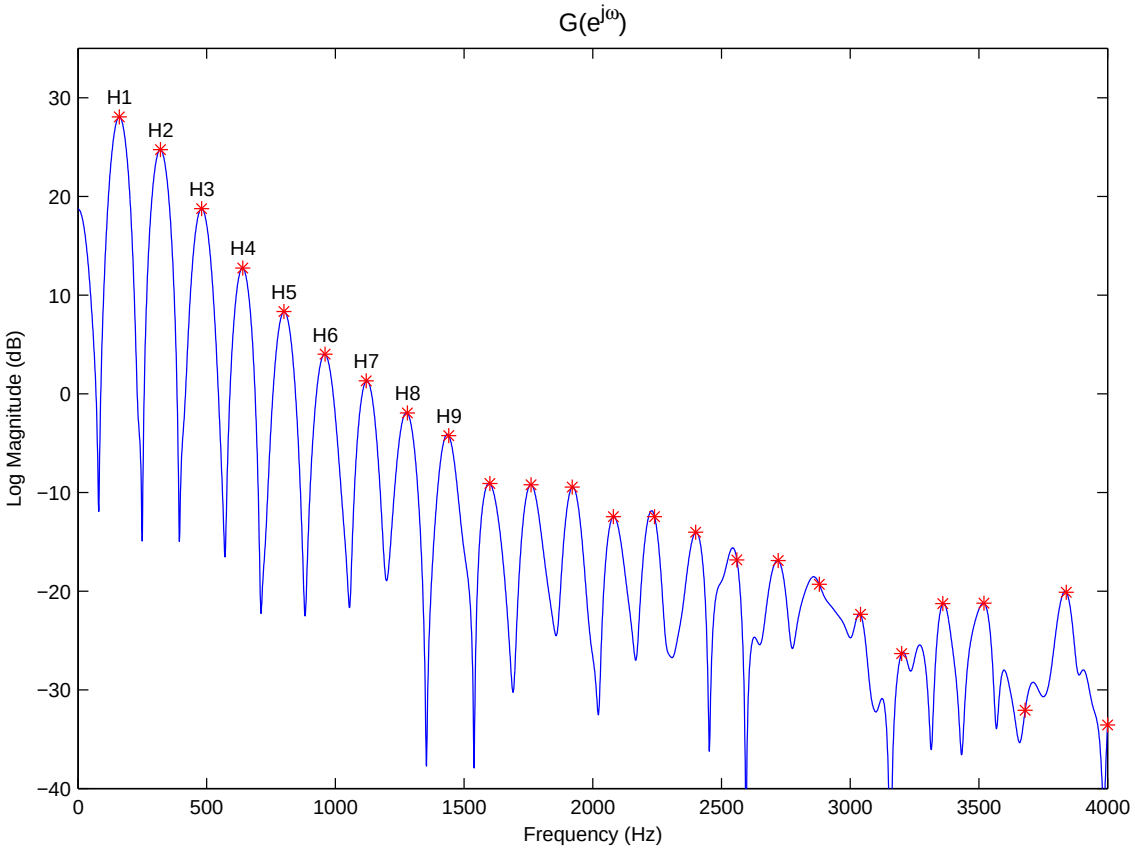


Figure 10: Magnitude spectrum of a glottal waveform estimate. Pitch harmonics are shown in red.

vocal tract with rapidly moving formants (diphthongs and semi-vowels)[93].

To construct an experimental setup amenable to inverse filtering, the majority of the analyses presented in this study focus on stationary vowels. Separately analyzing the speech for each stationary vowel contained in the corpus is problematic however, as the amount of data contained in the corpus for each vowel varies widely due to differences in the average length of phonation as well as differences in pronunciation among individual speakers. With the dual goal of using phonemes with a high number of observations while representing vocal tract spectral variations among vowels, three stationary vowels were chosen for this study: /iy/ (b<u>ee</u>t), /ae/ (b<u>a</u>t), and /ux/ (t<u>oo</u>t). These three vowels form a triangle in the F_1, F_2 plane [95] (where F_n denotes the center frequency of the n^{th} formant), with /iy/ having a low F_1 ($\approx$ 270 Hz) and a high F_2 ($\approx$ 2290 Hz), /ae/ a high F_1 ($\approx$ 660 Hz) and a moderate F_2 ($\approx$ 1720 Hz), and /ux/ a low F_1 ($\approx$ 300 Hz) and a low F_2 ($\approx$ 870 Hz) [95].

Feature extraction proceeded as follows: The speech data in the TIMIT corpus (Section 4.6) was divided into 25 ms frames with a frame step size of 10 ms. These are common frame length and step size values used speech recognition and speaker identification. Phonetic labels and time-alignment information were obtained from the transcription included with the TIMIT corpus. Each frame was assigned a phonetic label if it was found to lie completely within the boundaries of a single phoneme. Frames spanning two or more phonemes were discarded. The spectral envelope features described in Section 6.1 were computed for each frame. Also for each frame, the glottal waveform features given in Section 6.2 were estimated using each of the inverse filtering methods described in Chapter 5 (a model order of $p_{ap} = 16$ was used for all methods). Pitch information ($f_{0,k}$) was obtained for each frame k using the RAPT algorithm [109] implementation found in the Snack Sound Toolkit [102]. Glottal closure instants (GCIs) were also estimated for each frame, independently of the pitch estimates, using the DYPSA algorithm [85] as implemented in the VOICEBOX

toolbox [19]. Frames for which the average distance between GCIs differed from $1/f_{0,k}$ beyond a 20% tolerance value were deemed inconsistent and discarded.

In order to measure time-domain glottal waveform features, it is necessary for the speech frame to contain at least one complete, uninterrupted glottal cycle. Therefore, before performing inverse filtering, it was sometimes necessary to dynamically expand the frame size n_{wk} according to the following criteria:

$$n_{wk} = \begin{cases} 3 \,/\, f_{0,k} & \text{if } f_{0,k} < 120 \text{ Hz} \\ 25 \times 10^{-3} & \text{otherwise} \end{cases} \tag{41}$$

where n_{wk} is the length of the k^{th} frame, in milliseconds, and $f_{0,k}$ is its pitch, in Hz. The frames were expanded by adding an equal number of samples at the beginning and end of the frame, so that the frame centers would remain aligned with the centers of the 25 ms frames used to compute the spectral envelope features. On frames that contained two or more complete cycles, time-domain glottal feature values were averaged across the cycles within the frame in order to maintain a one-to-one observation correspondence with the spectral envelope features (i.e. one observation per frame). In the case of the FMIF glottal waveform estimation method, where the algorithm's output consisted of a vector of LF-model parameters for every pitch cycle, direct-measurement time- and frequency-domain parameters were computed from the synthesized LF-model waveforms.

6.3.1 Post-Processing

To facilitate the analysis of the extracted data, observations with extreme values were discarded using an outlier removal procedure based on order statistics. For a set of feature observations $\mathbf{X} = [\mathbf{x}_1 \, \mathbf{x}_2 \, ... \, \mathbf{x}_N]$, where $\mathbf{x}_i = [x_{1,i} \; x_{2,i} \; ... \; x_{M,i}]^T$ represents a single univariate ($M = 1$) or multivariate ($M > 1$) observation, a spread function

$$\mathcal{M}(\mathbf{x}_i) = \|\mathbf{x}_i - \tilde{\mathbf{x}}\| \tag{42}$$

Table 3: Minimum number of observations by dataset, phoneme, and gender.

Dataset	Male Speakers				Female Speakers			
	/iy/	/ae/	/ux/	Total	/iy/	/ae/	/ux/	Total
TRAIN1	21825	20977	7292	50094	11140	10528	3659	25327
TRAIN2	5663	5509	1766	12938	2944	2867	836	6647
TEST	9424	8689	2506	20619	6265	5160	1556	12981
Total:	36912	35175	11564	83651	20349	18555	6051	44955

was defined as the $\mathcal{L}^2$ distance from the median vector $\tilde{\mathbf{x}} = [\tilde{x}_1\, \tilde{x}_2 \ldots \tilde{x}_M]^T$. This function facilitated the detection of outliers on multivariate data, and was observed to be particularly useful for variables with two-tailed distributions with a long tail and a short tail. In this situation, the removal of observations in the top percentiles of $\mathcal{M}$ resulted in a heavier trimming of the long tail. In addition, for the removal of a small number of outliers, $\mathcal{M}$ was observed to behave well for single-tailed distributions and two-tailed distributions with similar tail lengths. Outlier removal was performed by removing observations above the 99^{th} percentile of $\mathcal{M}$, and was applied separately to each glottal feature and each SEF vector for the observations of each gender. Removed observations were tagged with a value of NaN, so that subsequent experiments involving an arbitrary combination of features could easily select the observations where none of the variables specific to the experiment had NaN values. Table 3 gives lower bounds on the number of available observations for each TIMIT subset, phoneme, and gender by listing the number of observations for which no single feature dimension contained an outlier.

6.4 Measurement Reliability of Glottal Waveform Features Obtained via Inverse Filtering

Before attempting to perform further analysis on a glottal waveform feature, it is prudent to first evaluate the feature's measurement reliability, particularly given the imprecise nature of inverse filtering. However, when only the acoustic speech signal

is available, there is no source of "ground truth" with which to compare the feature estimate, and an indirect assessment approach must be employed instead. Although it is understood that the characteristics of the glottal cycle are not entirely stationary even within small portions of a phoneme realization, it is also the case that independent noise arising from the feature estimation process (e.g., inverse filtering errors, model fitting errors) should, on average, result in increased distance between adjacent measurements of a glottal feature. Therefore, if it is assumed that the structure of the glottal cycle does not usually change abruptly during steady voicing, a measure of similarity between the features obtained from adjacent speech frames can be indicative of measurement reliability. This section describes the application of the above approach as a way to measure and compare the reliability of features computed from glottal waveform estimates obtained via inverse filtering.

For each IF method discussed in Chapter 5 and each glottal feature presented in Section 6.2.2, pairs of observations from voicing regions which can be assumed to be maximally "steady" were obtained from the TRAIN1 dataset by pairing a feature estimate u_i from the middle frame of each realization i of the three stationary vowels {/iy/, /ae/, /ux/} with the feature estimate v_i for the frame immediately following . To quantify the variation between frame pairs, the correlation coefficient $r_c(u, v)$ and the coefficient of determination $r_d(u, v)$ (Section 4.4) were computed, where $u = [u_1\, u_2\, ..\, u_N]$ and $v = [v_1\, v_2\, ...\, v_N]$ are the feature observations obtained from the center frames and the frames immediately following, respectively. Assuming that the variances of u and v are equal, r_d can be interpreted as a measure of the mean-squared error (MSE) between adjacent estimates, expressed as a fraction of the variance of u or v (Equation 8). A value of $r_d = 1$ implies that the MSE is zero ($u = v$) and values below zero indicate that the MSE between u and v is actually greater than the variance of the feature's distribution over the entire phoneme. The results for each gender, phoneme, and inverse filtering method are given in Figures 11–14.

The results show wide variability across features and IF methods, with correlation coefficient values below 0.2 and above 0.9. Table 4 shows the mean r_d values (across the three phonemes) of the best-performing IF method for each feature. The females show much higher overall values of r_d than the males. This difference seems to be mostly due to the more extreme lower-bound of r_d found in the males, as both the male and female speakers produced r_d values of ≈ 0.76 for their "most reliable" glottal feature. For the female speakers, the $IAIF$ method consistently produced the most reliable features almost without exception. For the males, the IF method with the highest reliability varied with respect to the glottal feature, mostly alternating between closed-phase inverse filtering ($CPIF$) and $IAIF$.

The overall advantage of $IAIF$ on female speakers is consistent with the characteristics of female glottal cycles, which are more likely to contain an incomplete closed glottal phase [69, 103, 62, 73] or one that is too short to fit an LP analysis window, thus hindering the performance of the closed-phase inverse filtering method. The global mean values of r_d for males (0.149) and females (0.4250) illustrate how error-prone glottal feature extraction via IF can be, as values of $r_d < 0.5$ indicate that the MSE between adjacent measurements is more than 0.5 times the overall variance of the feature across the data. This assertion is further supported by the numerous feature-IF method combinations for which $r_d < 0$, as shown in Figures 13–14, and by the excellent measurement reliability results for the pitch feature (f_0), shown in Table 5. The f_0 feature, which obtained r_c and r_d values above 0.99 and 0.98, respectively, can be regarded as an example of the high values that can be obtained by the reliability measurement method under discussion for a glottal feature that can be easily estimated directly from the speech signal with a known low error rate.

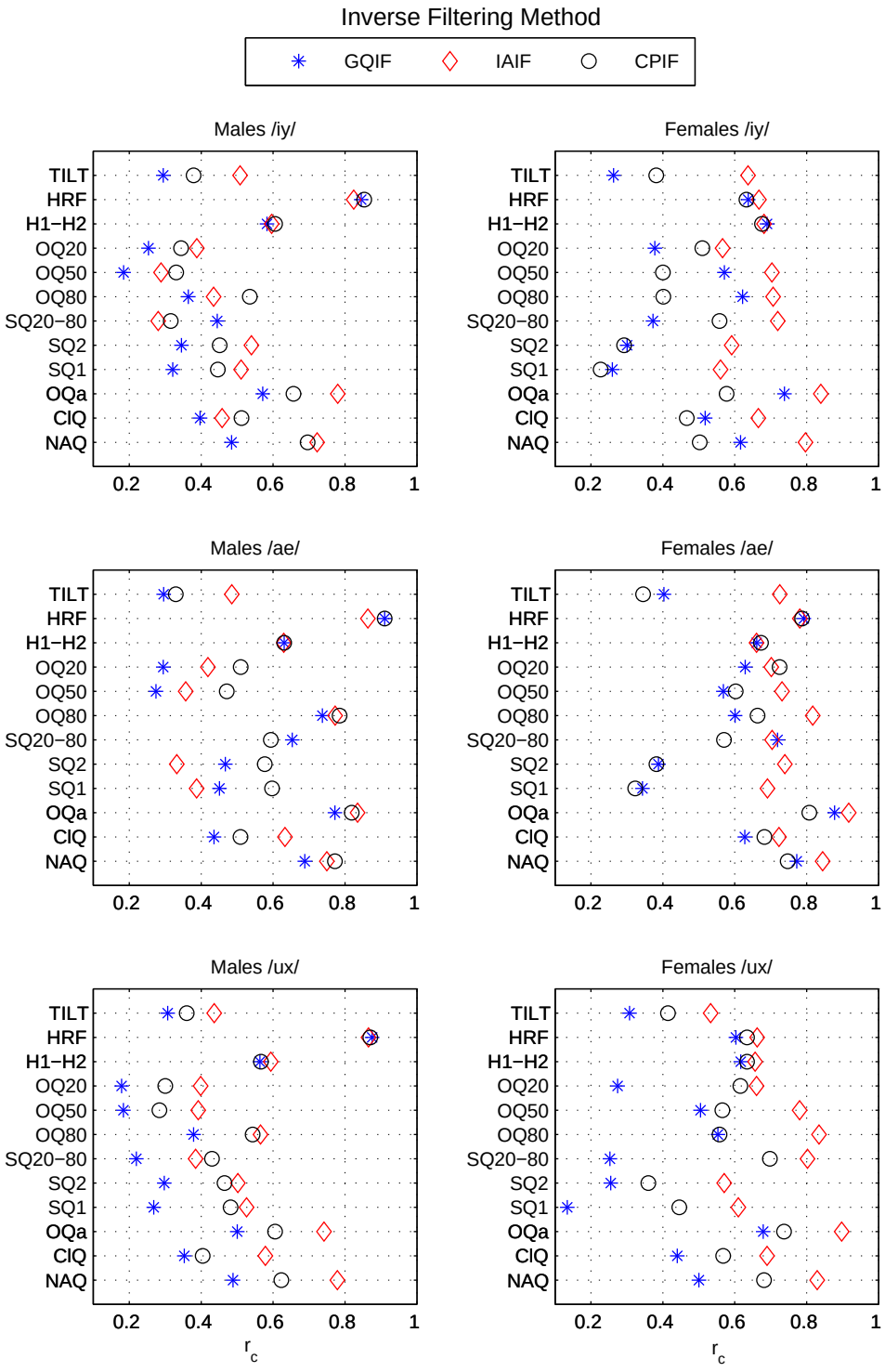


Figure 11: Measurement reliability: correlation coefficient r_c between observation pairs from adjacent frames, direct measurement features. Male and female speakers, stationary vowels /ae/, /iy/, /ux/.

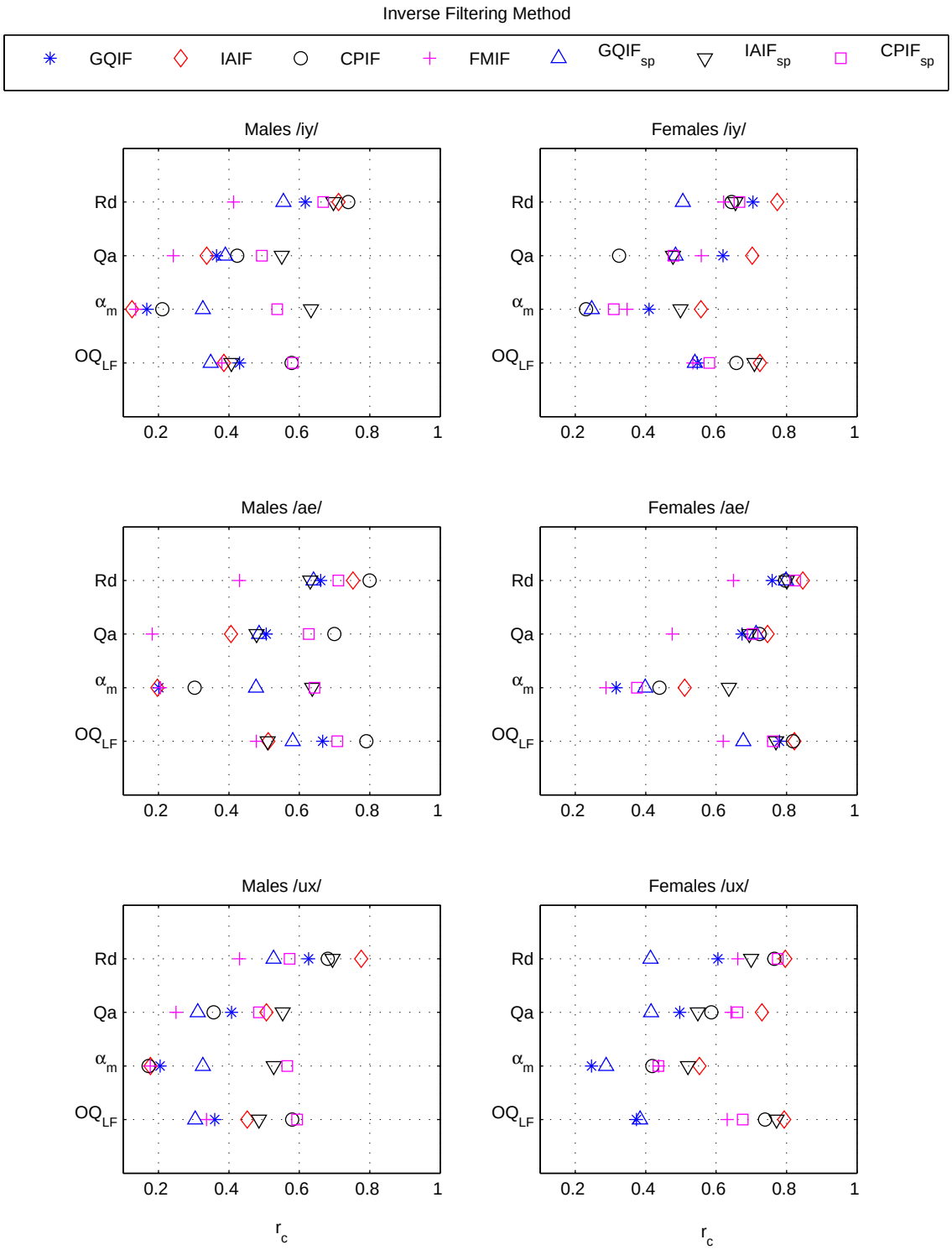


Figure 12: Measurement reliability: correlation coefficient r_c between observation pairs from adjacent frames, LF-model features. Male and female speakers, stationary vowels /ae/, /iy/, /ux/. The *sp* subscript denotes frequency-domain LF model fitting (Section 6.2.2.2).

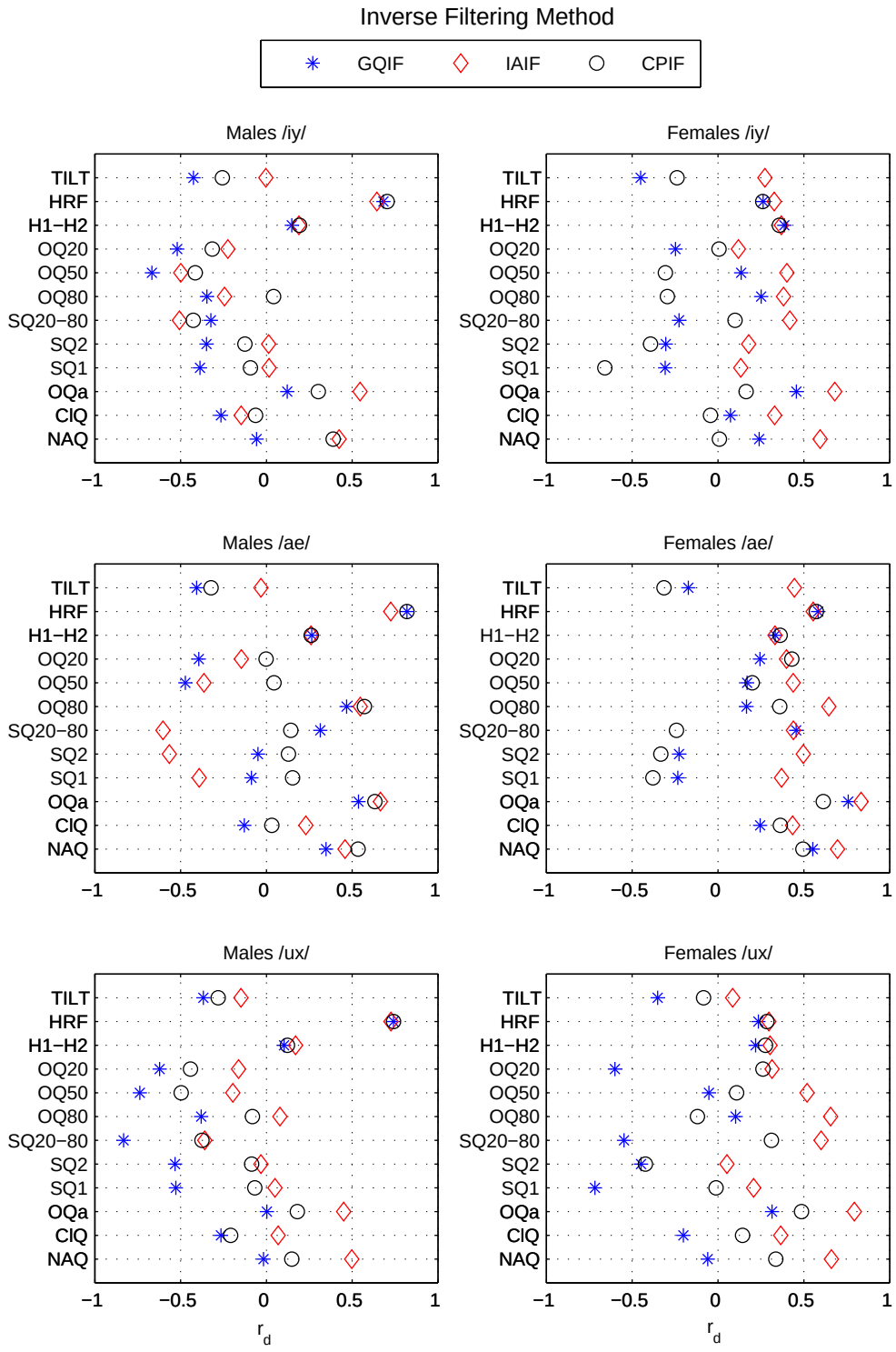


Figure 13: Measurement reliability: coefficient of determination r_d between observation pairs from adjacent frames, direct measurement features. Male and female speakers, stationary vowels /ae/, /iy/, /ux/.

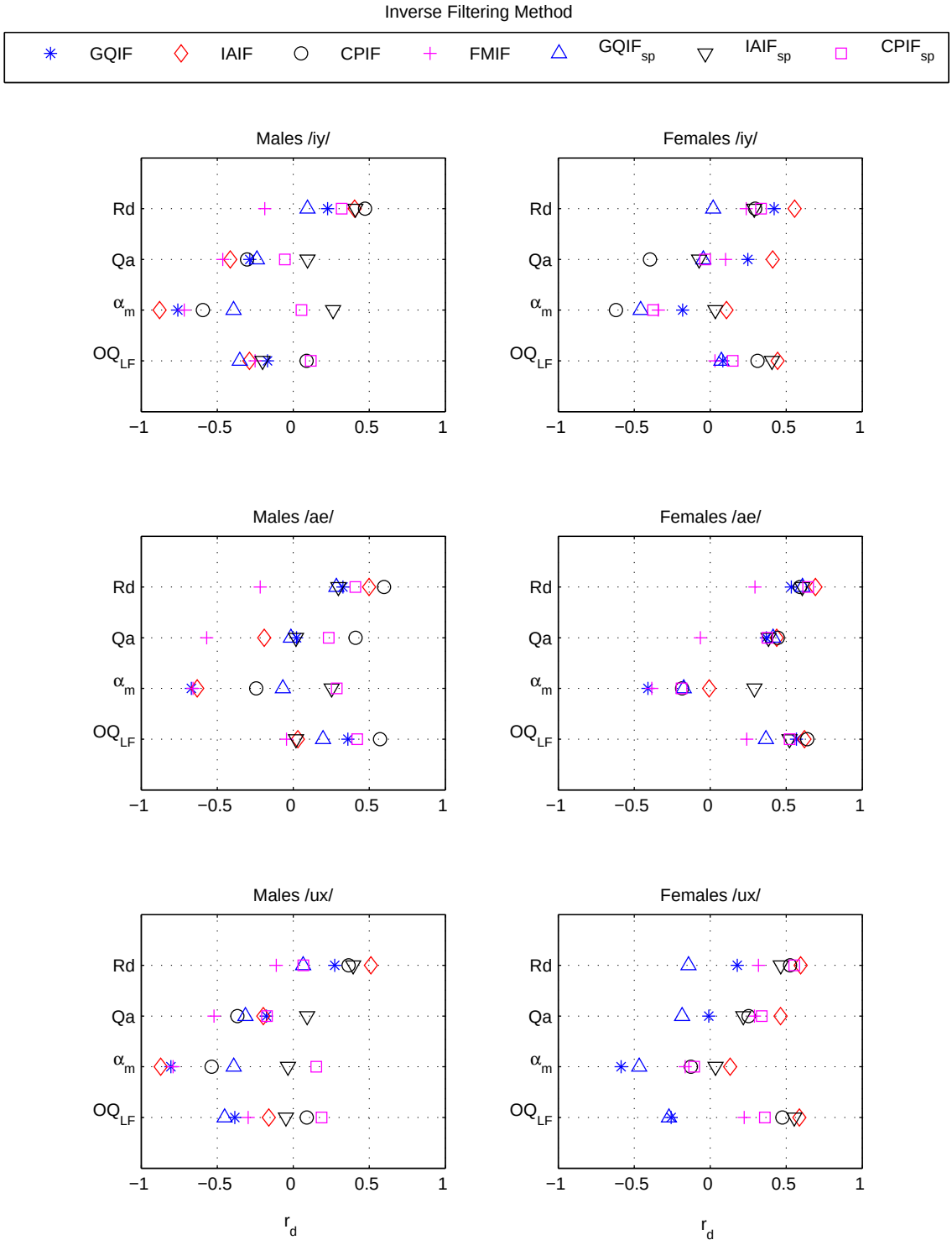


Figure 14: Measurement reliability: coefficient of determination r_d between observation pairs from adjacent frames, LF-model features. Male and female speakers, stationary vowels /ae/, /iy/, /ux/. The *sp* subscript denotes frequency-domain LF model fitting (Section 6.2.2.2).

Table 4: Inverse filtering methods with highest measurement reliability across phonemes

Males			Females		
Feature	Best Method	$\bar{r}_d$	Feature	Best Method	$\bar{r}_d$
HRF	$CPIF$	0.755	OQa	$IAIF$	0.769
OQ_a	$IAIF$	0.554	NAQ	$IAIF$	0.650
Rd	$CPIF$	0.478	Rd	$IAIF$	0.614
NAQ	$IAIF$	0.460	OQ_{80}	$IAIF$	0.561
OQ_{LF}	$CPIF$	0.249	OQ_{LF}	$IAIF$	0.550
$H1$-$H2$	$IAIF$	0.207	$SQ_{20\text{-}80}$	$IAIF$	0.486
OQ_{80}	$CPIF$	0.178	OQ_{50}	$IAIF$	0.453
α_m	$CPIF_{sp}$	0.164	Qa	$IAIF$	0.436
Qa	$IAIF_{sp}$	0.067	HRF	$IAIF$	0.393
ClQ	$IAIF$	0.050	ClQ	$IAIF$	0.377
SQ_1	$CPIF$	-0.003	$H1$-$H2$	$IAIF$	0.335
SQ_2	$CPIF$	-0.027	OQ_{20}	$IAIF$	0.277
$TILT$	$IAIF$	-0.061	$TILT$	$IAIF$	0.268
OQ_{20}	$IAIF$	-0.177	SQ_2	$IAIF$	0.242
$SQ_{20\text{-}80}$	$CPIF$	-0.220	SQ_1	$IAIF$	0.237
OQ_{50}	$CPIF$	-0.289	α_m	$IAIF_{sp}$	0.120
Mean:		0.1490	Mean:		0.4230

Table 5: Measurement reliability for pitch (f_0) estimates obtained via the RAPT algorithm. Male and female speakers, stationary vowels /ae/, /iy/, /ux/.

	Males			Females		
	/iy/	/ae/	/ux/	/iy/	/ae/	/ux/
r_c	0.992	0.994	0.988	0.991	0.993	0.992
r_d	0.983	0.987	0.975	0.982	0.985	0.982

6.4.1 Glottal Waveform Feature Statistics

The mean and standard deviations for each glottal waveform feature, after outlier removal, are given in Table 6 for each gender and phoneme. Each feature was obtained using the IF method showing the highest measurement reliability, as listed in Table 4. Most features show gender differences that are in general agreement with known properties of the glottal source for male and female speakers. Apart from the obviously higher mean pitch for female speakers, which is accompanied by a higher standard deviation due to the logarithmic scale of pitch perception, female speakers obtained higher values of the open quotient, which is consistent with the known existence of incomplete closed phases of female speakers [103, 62, 73].

The higher open quotient of female speakers was accompanied by lower values of $H1$-$H2$, which is contrary to the expected decrease in the center frequency of the glottal formant due to the higher OQ, and to previous findings of higher $H1$-$H2$ values for female speakers [69, 56]. It should be noted, however, that those studies involved speech material where sustained vowels had been deliberately prompted, so that the difference in results may be due to the increased diversity of text materials and speakers in the larger TIMIT database. This possibility is supported by the relatively large intra-gender variation for the $H1$-$H2$ feature (approx. 7–8 dB standard deviation), which is about the same as the average difference between gender means (7.12 dB). In addition, even in the theoretical LF-model of the glottal waveform, $H1$-$H2$ has been shown to be jointly related not only to the open quotient but also to the speed and return quotients [40, 36].

Incomplete glottal closure in females speakers is also consistent with the observed increases in the return quotient (Qa), closing quotient (ClQ) and normalized amplitude quotient (NAQ), which are all indicative of slower and more gradual vocal fold adduction. An interesting result was the decrease in spectral tilt (less negative values indicate a decreased spectral slope) for female speakers, which is inconsistent with

Table 6: Mean value and standard deviation (shown in parentheses) for each glottal feature, computed on the TRAIN1 dataset from glottal waveforms obtained via the IF methods in Table 4.

		/iy/		/ae/		/ux/	
f_0	Males	122	(25)	118	(24)	125	(25)
(Hz)	Females	212	(44)	197	(45)	208	(42)
OQ_{LF}	Males	0.383	(0.11)	0.356	(0.090)	0.373	(0.11)
	Females	0.461	(0.14)	0.374	(0.11)	0.464	(0.15)
α_m	Males	0.809	(0.020)	0.810	(0.018)	0.810	(0.019)
	Females	0.828	(0.017)	0.825	(0.015)	0.826	(0.017)
Qa	Males	0.0401	(0.052)	0.0232	(0.026)	0.0363	(0.045)
	Females	0.0923	(0.080)	0.0961	(0.075)	0.0860	(0.067)
Rd	Males	0.466	(0.20)	0.430	(0.18)	0.475	(0.19)
	Females	1.64	(0.65)	1.29	(0.56)	1.66	(0.65)
NAQ	Males	0.0665	(0.030)	0.0582	(0.025)	0.0722	(0.030)
	Females	0.159	(0.068)	0.123	(0.056)	0.168	(0.070)
ClQ	Males	0.170	(0.097)	0.145	(0.075)	0.183	(0.083)
	Females	0.340	(0.14)	0.269	(0.12)	0.356	(0.14)
OQa	Males	0.250	(0.066)	0.223	(0.058)	0.271	(0.066)
	Females	0.398	(0.12)	0.320	(0.097)	0.403	(0.12)
SQ_1	Males	3.28	(1.9)	2.81	(1.4)	3.33	(2.0)
	Females	1.36	(0.73)	1.36	(0.60)	1.26	(0.64)
SQ_2	Males	1.79	(1.1)	1.49	(0.83)	1.88	(1.2)
	Females	0.690	(0.48)	0.733	(0.41)	0.628	(0.43)
$SQ_{20\text{-}80}$	Males	0.469	(0.51)	0.392	(0.32)	0.465	(0.52)
	Females	0.795	(0.69)	0.512	(0.47)	0.784	(0.65)
OQ_{80}	Males	0.0753	(0.043)	0.0695	(0.033)	0.0750	(0.045)
	Females	0.106	(0.055)	0.0817	(0.037)	0.111	(0.062)
OQ_{50}	Males	0.202	(0.12)	0.171	(0.082)	0.205	(0.13)
	Females	0.273	(0.10)	0.205	(0.074)	0.275	(0.10)
OQ_{20}	Males	0.652	(0.19)	0.666	(0.19)	0.637	(0.17)
	Females	0.562	(0.17)	0.549	(0.18)	0.554	(0.17)
$H1$-$H2$	Males	1.89	(8.5)	3.13	(8.0)	1.44	(8.2)
(dB)	Females	-4.56	(7.3)	-5.53	(7.8)	-4.82	(7.2)
HRF	Males	9.81	(6.7)	8.09	(5.9)	10.8	(6.4)
	Females	29.7	(7.3)	29.3	(8.4)	28.9	(7.2)
$TILT$	Males	-29.0	(4.2)	-29.3	(4.2)	-29.2	(4.1)
(dB/dec)	Females	-21.7	(4.7)	-21.9	(4.7)	-22.5	(4.3)

the results in [69], where a more abducted glottal configuration is shown to be related to an increase in spectral roll-off. A possible cause for this result is the aspiration noise that is associated with incomplete glottal closure, which raises the noise floor of the glottal spectrum, thus potentially decreasing the measurement of spectral slope if this measurement includes a frequency interval where the aspiration noise is higher in amplitude than the pitch harmonics. Such an interval is likely included in the $TILT$ feature, as it goes all the way to the Nyquist frequency (8 kHz), where little harmonic energy is expected. Aspiration noise, in combination with more closely-spaced harmonics due to higher pitch, also explain the much higher value of HRF for female speakers.

For most glottal features, between-phoneme variation was much lower than within-phoneme standard deviation, suggesting that these features are largely unaffected by inter-phoneme differences in vocal tract configuration. Other features, however, showed noticeably lower values for /ae/ than for the other vowels. These included Qa (males only); OQa, NAQ, ClQ (both genders); OQ_{20} and $SQ_{20\text{-}80}$ (females only). The inter-phoneme differences, which may be due to how the high first-formant of the /ae/ vowel affects the inverse filtering process and the resulting glottal waveform (/iy/ and /ux/ tend to have a similarly low first-formant), highlight the importance of considering multiple methods (e.g., time-domain v.s. frequency-domain, threshold-based v.s. amplitude-based) for measuring the same underlying properties of the glottal source.

Since the 16 glottal features chosen for this study span several ways to measure a smaller set of salient properties of the glottal waveform, the Spearman rank correlation coefficient r_r between features was computed for male and female speakers (Tables 7 and 8) to detect pairwise linear and non-linear monotonic relationships. Some interesting and unexpected results are worth mentioning. It should first be emphasized, however, that the absence of a strong rank correlation between a pair of

features does not necessarily indicate that the features are independent or unrelated. Such a result could arise from a more complicated non-linear relationship, or due to a multivariate relationship where the first feature in the pair jointly depends on a group of features that includes second feature.

For male speakers, the highest correlation coefficient was between NAQ and ClQ (0.78), showing that the normalized amplitude quotient did indeed approximate the closing quotient, as intended. Interestingly, OQa was moderately correlated with NAQ and ClQ (0.58 and 0.51, respectively), but not strongly correlated with any other measure of open quotient, suggesting that this feature may be measuring something other than open quotient. Similarly, $SQ_{20\text{-}80}$ was strongly correlated with OQ_{LF} (0.68) but not with any measure of speed quotient, while OQ_{LF} did not appear to be strongly correlated to any other OQ measure. The OQ_{20} feature was not strongly correlated with any other feature, indicating that the 20% amplitude threshold may not be a useful method for measuring open quotient. Two of the threshold-based open quotients (OQ_{80} and OQ_{50}) did show an expected correlation (0.62), as did the primary and secondary speed quotients SQ_1 and SQ_2 (0.64). The harmonic richness factor (HRF) showed some correlation with spectral tilt (0.51) and a strong correlation to f_0 (0.76). Both of these relationships are expected due to the definition of HRF.

Spectral tilt appeared to be uncorrelated with the return phase quotient Qa (0.03), which would contradict theoretical results on the spectra of the LF model [40, 36]. This may be due to the extended frequency range over which $TILT$ was computed, which may be causing this feature to be primarily a measure of aspiration noise. Another result that contradicts the analysis of LF waveforms and their spectra is the near zero correlation between Rd and $H1$-$H2$ (0.08), as these features were shown to be linearly related in [40]. In general, apart from the correlation between Rd and

f_0, LF-model features did not show much correlation with any other feature, suggesting that the direct-measurement and LF-model fitting methods of glottal waveform measurement may yield a fundamentally different set of features.

The pairwise feature correlations for female speakers were generally higher than for males, but followed a similar pattern. Notable exceptions were a decreased correlation between HRF and f_0, as well as a strong correlation between OQa and OQ_{LF}. The amplitude-based open quotient (OQa) obtained moderate to strong correlation with $SQ_{20\text{-}80}$ and the open quotients OQ_{80} and OQ_{50} (0.63, 0.68, and 0.69, respectively). This is in addition to the strong correlation with the closing quotient measures ClQ and NAQ. Unlike the case of male speakers, ClQ showed a strong negative correlation to the primary and secondary speed quotients (-0.74, and -0.66, respectively), while Rd appeared to be correlated to NAQ, ClQ, and OQa. From these results, it appears that OQa is behaving as a function of both open and closing quotients, while Rd holds a close relationship to the closing quotient that it did not hold for the male speakers. In addition, the negative correlation between ClQ and the speed quotients suggests that for female speakers, variations in the asymmetry of the open glottal phase may be primarily due to a change in the length of the closing phase, with the length of the opening phase remaining relatively unchanged.

Table 7: Spearman rank correlation coefficient r_r between pairs of glottal features, averaged across phonemes. Computed on the TRAIN1 dataset from glottal waveforms obtained via the IF methods in Table 4. Male speakers.

	f_0	OQ_{LF}	α_m	Qa	Rd	NAQ	ClQ	OQa	SQ_1	SQ_2	SQ_{20-80}	OQ_{80}	OQ_{50}	OQ_{20}	$H1$-$H2$	HRF	$TILT$
f_0	1.00	-.04	.12	.25	.62	.4	.31	.28	-.37	-.27	.01	.15	.12	-.15	.1	**.76**	.46
OQ_{LF}	-.04	1.00	-.18	.03	.37	.15	.13	.32	.36	.05	.68	.3	.39	-.04	.05	-.24	-.04
α_m	.12	-.18	1.00	.25	-.01	.23	.17	-.06	-.37	-.38	-.34	-.04	-.21	0	-.05	.21	0
Qa	.25	.03	.25	1.00	.26	.36	.27	.19	-.3	-.23	-.03	.21	.13	-.06	0	.14	.03
Rd	.62	.37	-.01	.26	1.00	.45	.37	.39	-.27	-.23	.23	.37	.42	-.09	.08	.33	.27
NAQ	.4	.15	.23	.36	.45	1.00	**.78**	.58	-.38	-.32	.11	.22	.2	-.19	.04	.21	-.02
ClQ	.31	.13	.17	.27	.37	**.78**	1.00	.51	-.33	-.3	.07	.23	.17	-.16	.03	.15	-.03
OQa	.28	.32	-.06	.19	.39	.58	.51	1.00	-.12	.04	.32	.42	.4	-.13	.05	.01	-.04
SQ_1	-.37	.36	-.37	-.3	-.27	-.38	-.33	-.12	1.00	.64	.33	-.22	-.01	.04	-.03	-.2	-.11
SQ_2	-.27	.05	-.38	-.23	-.23	-.32	-.3	.04	.64	1.00	.12	.03	.19	.11	.01	-.2	-.1
$SQ_{20\text{-}80}$	.01	.68	-.34	-.03	.23	.11	.07	.32	.33	.12	1.00	.18	.36	-.14	.08	-.2	0
OQ_{80}	.15	.3	-.04	.21	.37	.22	.23	.42	-.22	.03	.18	1.00	.62	.13	.08	-.17	-.04
OQ_{50}	.12	.39	-.21	.13	.42	.2	.17	.4	-.01	.19	.36	.62	1.00	.11	.07	-.13	.01
OQ_{20}	-.15	-.04	0	-.06	-.09	-.19	-.16	-.13	.04	.11	-.14	.13	.11	1.00	-.02	-.13	-.07
$H1$-$H2$	.1	.05	-.05	0	.08	.04	.03	.05	-.03	.01	.08	.08	.07	-.02	1.00	-.04	-.03
HRF	**.76**	-.24	.21	.14	.33	.21	.15	.01	-.2	-.2	-.2	-.17	-.13	-.13	-.04	1.00	.51
$TILT$	.46	-.04	0	.03	.27	-.02	-.03	-.04	-.11	-.1	0	-.04	.01	-.07	-.03	.51	1.00

Table 8: Spearman rank correlation coefficient r_r between pairs of glottal features, averaged across phonemes. Computed on the TRAIN1 dataset from glottal waveforms obtained via the IF methods in Table 4. Female speakers.

	f_0	OQ_{LF}	α_m	Qa	Rd	NAQ	ClQ	OQa	SQ_1	SQ_2	SQ_{20-80}	OQ_{80}	OQ_{50}	OQ_{20}	$H1$-$H2$	HRF	$TILT$
f_0	1.00	.38	.12	.4	.59	.48	.43	.49	-.19	-.25	.33	.4	.49	.11	.5	.51	.54
OQ_{LF}	.38	1.00	.11	.27	.36	.44	.35	.65	.23	-.17	**.71**	.6	.56	.26	.15	.07	.05
α_m	.12	.11	1.00	.19	.07	.07	.04	.14	.02	.06	.05	.15	.11	.01	.02	.1	.04
Qa	.4	.27	.19	1.00	.35	.39	.32	.47	-.17	-.09	.28	.45	.4	0	.17	.14	.12
Rd	.59	.36	.07	.35	1.00	**.77**	.68	.69	-.5	-.49	.38	.43	.58	-.12	.25	.22	.19
NAQ	.48	.44	.07	.39	**.77**	1.00	**.82**	**.76**	-.56	-.57	.36	.56	.62	-.01	.18	.12	.05
ClQ	.43	.35	.04	.32	.68	**.82**	1.00	.64	**-.74**	-.66	.28	.44	.53	-.03	.17	.1	.07
OQa	.49	.65	.14	.47	.69	**.76**	.64	1.00	-.27	-.25	.63	.68	.69	.04	.18	.13	.04
SQ_1	-.19	.23	.02	-.17	-.5	-.56	**-.74**	-.27	1.00	.64	.15	-.11	-.18	.25	-.04	-.07	-.03
SQ_2	-.25	-.17	.06	-.09	-.49	-.57	-.66	-.25	.64	1.00	-.12	-.1	-.15	.16	-.08	-.11	-.09
$SQ_{20\text{-}80}$	.33	**.71**	.05	.28	.38	.36	.28	.63	.15	-.12	1.00	.49	.44	.01	.13	.05	.04
OQ_{80}	.4	.6	.15	.45	.43	.56	.44	.68	-.11	-.1	.49	1.00	**.78**	.2	.15	.05	-.01
OQ_{50}	.49	.56	.11	.4	.58	.62	.53	.69	-.18	-.15	.44	**.78**	1.00	.23	.22	.07	.06
OQ_{20}	.11	.26	.01	0	-.12	-.01	-.03	.04	.25	.16	.01	.2	.23	1.00	.11	.02	.04
$H1$-$H2$	.5	.15	.02	.17	.25	.18	.17	.18	-.04	-.08	.13	.15	.22	.11	1.00	.02	.29
HRF	.51	.07	.1	.14	.22	.12	.1	.13	-.07	-.11	.05	.05	.07	.02	.02	1.00	.68
$TILT$	.54	.05	.04	.12	.19	.05	.07	.04	-.03	-.09	.04	-.01	.06	.04	.29	.68	1.00

CHAPTER VII

TRANSFORMATION OF SPECTRAL ENVELOPE FEATURES INTO GLOTTAL WAVEFORM FEATURES

The proposed method of glottal feature extraction is motivated by the limitations associated with existing inverse filtering methods and by the fact that in many applications it is the *features* of the glottal waveform, and not the exact waveform itself, that are of interest. Given that the exact shape or spectrum of the glottal waveform is difficult to separate from the vocal tract resonances of the acoustic speech signal, this chapter explores the possibility of obtaining, by learning statistical transformations of the spectral envelope, equally useful and/or more robust glottal features than those obtainable through inverse filtering. To this end, each of the four spectral envelope feature vectors (SEFs) described in Section 6.1 were augmented with each of the inverse filtering (IF) glottal waveform features listed in Section 6.2 and used to train Gaussian mixture models (GMMs) to learn their joint probability density functions (PDFs), as described in Section 4.2. These GMMs were then used to estimate, from a separate set of observations, the glottal waveform features using only the SEFs as input, according to Equation 5.

The success of each model in transforming SEFs into glottal features via Gaussian mixture regression (GMR) was evaluated in three ways: First, the correlation coefficient r_c and coefficient of determination r_d between the GMR and inverse filtering features were used to quantify how closely the GMM regression can match the features that it was trained to reproduce. These measures were then used to select the most useful SEF for glottal feature estimation (Section 7.1) and to determine whether the use of pitch information (f_0) in addition to SEFs significantly improved the GMR

feature estimates (Section 7.2). Next, in Section 7.3, the correlation coefficient and coefficient of determination between adjacent observations were used to quantify the measurement reliability of the GMR glottal features, allowing for direct comparison to the reliability results for IF glottal features. Finally, the merits of GMR and IF glottal features in a speech analysis application were evaluated and compared through a series of pairwise speaker identification experiments (Section 7.5) using single glottal features, a combination of glottal features, and SEFs combined with glottal features.

7.1 Evaluation of Spectral Envelope Feature Vectors for Glottal Feature Estimation

As stated in Section 6.1, Spectral envelope features were designed to discard spectral details in order to produce a lower-dimensional representation of speech that captures the important broad attributes of the magnitude spectrum. Commonly, what is defined as "important" is the set of formants and anti-formants that carry much of the phonetic information. The intention is then to minimize the contribution of other sources of variability in the speech spectrum, which include fundamental frequency and shape of the glottal waveform. However, changes in the glottal waveform obviously affect its spectrum, and it has been shown, at least in the case of synthetic glottal waveform signals, that changes in glottal time-domain features affect the glottal waveform spectrum in a broad manner, by altering the location and width of a wide formant and the spectral roll-off at high frequencies [40, 12, 36]. These changes to the glottal spectral *envelope* will inevitably affect the spectral envelope of speech as well, and this is part of the reason why the glottal source plays an important perceptual role in voice quality [48] and voice identity [22, 68, 67]. However, different SEFs are computed through varying amounts of preprocessing and may operate of somewhat different perceptual frequency scales. As such, the extent to which glottal information will manifest itself on a particular SEF set is likely to vary among them. This section describes a procedure designed to measure this variability in order to

find, among the SEFs described in Section 6.1, the one that is most adequate for estimating glottal waveform features via GMM-regression.

7.1.1 Training Procedure

The observations of SEFs $\mathbf{x}_{i,m}$ and glottal features $y_{j,m}$, where m is the frame index, $i \in \{mfcc, plp, melsub, melsub_{41}\}$ denotes the spectral feature vector, and

$$j \in \left\{ \begin{array}{l} OQ_{lf},\, \alpha_m,\, Qa,\, Rd,\, NAQ,\, ClQ,\, OQ_a,\, SQ_1,\, SQ_2, \\ SQ_{2080},\, OQ_{80},\, OQ_{50},\, OQ_{20},\, H1\text{-}H2,\, HRF,\, TILT \end{array} \right\}$$

indicates a particular glottal waveform feature, were obtained from the TRAIN1 dataset (Section 4.6). Each SEF vector was augmented with a glottal feature, producing the vectors

$$\mathbf{w}_{i,j,m} = \left[\mathbf{x}_{i,m}^T\ y_{j,m}\right]^T.$$

For each phoneme $phn \in \{/iy/, /ae/, /ux/\}$ and gender $g \in \{male, female\}$, GMM representations $f_{i,j}$ of the distributions $\mathbf{w}_{i,j}$ were trained via the EM algorithm, as described in Section 4.2. The IF method was chosen separately for each glottal feature and for each gender as the one with the highest average measurement reliability across all phn, as shown in Table 4. The stopping criterion for the EM algorithm consisted of a threshold t_{EM} on the improvement of the log-likelihood function $\frac{1}{M}\sum_{m=1}^{M}\log(f_{i,j}(\mathbf{w}_{i,j,m}))$ over consecutive iterations. Training stopped when t_{EM} was less than 1×10^{-4}. To minimize the effect of convergence to global minima, three training sessions were performed per GMM, and the one with the highest final likelihood was kept. Separate GMMs were trained using 2, 4, 8, and 16 Gaussian components, as larger models would have required more than the limited number of frames available for single phonemes (Table 3).

7.1.2 Evaluation

Testing was carried out on the TRAIN2 subset of TIMIT, which contains a set of speakers independent from those in the TRAIN1 subset (Section 4.6). For each frame

k of the testing data, the glottal features were estimated from the SEFs as

$$\hat{y}_{j,k} = \mathcal{F}_{f_{i,j}}(\mathbf{x}_{i,k}),$$

where $\mathcal{F}_f(\mathbf{x})$ is the GMM regression (GMR) function defined by Equation 5. The accuracy of the GMM regression procedure was measured by computing the correlation coefficient $r_c(y_j, \hat{y}_j)$ and the coefficient of determination $r_d(y_j, \hat{y}_j)$ between the glottal features obtained via IF and those obtained via GMR.

The correlation coefficient r_c quantifies the degree to which y_j and $\hat{y}_j$ approximate each other, up to an arbitrary scale and shift operation. The interpretation of r_d in this context is as a measure of the mean-squared error (MSE) between $\hat{y}_j$ and y_j, expressed as a fraction of the variance in the IF-derived feature observations y_j (Equation 8). A value of $r_d(y_j, \hat{y}_j) = 0$ represents what would be obtained using a 0^{th} order estimator $\hat{y}_{j,k} = \frac{1}{K}\sum_{m=1}^{K} y_{j,k}$, which does not depend on $\mathbf{x}_{j,k}$ and simply outputs the mean value of the glottal waveform feature over the training data.

7.1.3 Results

The resulting values of r_c and r_d for each combination of SEF and glottal feature are given in Appendix A.1, while r_c and r_d averages across all glottal features are listed by SEF in Tables 9 and 10, and by number of GMM components (N_r) in Tables 11 and 12. From Table 10, it can be seen that the use of *melsub* features consistently results in GMR features that more closely match their IF counterparts when compared to $mfcc$ or *plp* SEFs. The increase in spectral resolution afforded by the 41-subband $melsub_{41}$ SEF vector results in an additional increase in r_d, particularly for males. A possible explanation for this gender-specific result is that for the lower-pitched male speakers, the glottal waveform harmonics are more closely spaced than for the female speakers, thus requiring somewhat finer spectral resolution to become individually identifiable. When compared across phonemes, average r_d values are consistently lower for /ux/ than for /iy/ and /ae/, which is most likely due to

the lower number of training observations available for this vowel (Table 3).

From Table 12, it can be seen that the use of 4-component GMMs results in the highest mean r_d for both genders. The results again suggest a limitation due to the finite number of observations (Table 3), with /ux/ favoring smaller GMMs more so than /iy/ and /ae/. While the results in Table 11 do show additional improvement of r_c for 8-component GMMs, the r_d measure is preferred for the selection of an appropriate model size N_r, since this measure is affected not only by the linear relationship between the estimated and target values but also by bias and scale differences (Section 4.4).

Tables 13 and 14 list the average correlation coefficient and coefficient of determination, respectively, for each feature, obtained using 4-component GMMs trained on $melsub_{41}$ features. From these tables it is evident that the accuracy of GMM glottal feature estimation varied widely with respect to the specific glottal feature. The average correlation coefficient was as high as 0.855 (HRF, males) and as low as 0.108 (SQ_2, females). Two features for each gender obtained average values of $\bar{r}_c > 0.70$ and $\bar{r}_d > 0.49$ (Rd, and HRF for males; HRF, and $H1$-$H2$ for females). As a baseline for comparison, the correlation and determination coefficients between pitch estimates obtained directly from the speech signal via (RAPT) and those estimates from the $melsub_{41}$ SEF via GMR are given in Table 15. The obtained values for pitch were substantially higher, with correlation coefficients well above .80 and as high as 0.93. This result is consistent with a recent study that has shown success in estimating pitch from MFCCs [76].

Three male features (SQ_2, SQ_{20-80}, OQ_{20}) and four female features (α_m, SQ_1, SQ_2, OQ_{20}) obtained values of r_d near or below zero for every phoneme, which could suggest that for these features GMM regression performed no better than a simple estimator that always outputs the mean value of the feature over the training data. However, these features show a correlation coefficient as high as 0.435 in Table 13,

Table 9: Mean correlation coefficient r_c between IF and GMR glottal features, by spectral envelope feature set.

	Males				Females			
Feature Set:	/iy/	/ae/	/ux/	Mean:	/iy/	/ae/	/ux/	Mean:
$melsub_{41}$	0.512	0.560	0.482	0.518	0.513	0.550	0.435	0.499
$melsub$	0.409	0.470	0.344	0.407	0.484	0.477	0.399	0.453
$mfcc$	0.379	0.434	0.318	0.377	0.361	0.364	0.265	0.330
plp	0.284	0.358	0.217	0.286	0.315	0.293	0.200	0.269
Mean:	0.396	0.455	0.340	0.397	0.418	0.421	0.324	0.388

Table 10: Mean coefficient of determination r_d between IF and GMR glottal features, by spectral envelope feature set.

	Males				Females			
Feature Set:	/iy/	/ae/	/ux/	Mean:	/iy/	/ae/	/ux/	Mean:
$melsub_{41}$	0.273	0.321	0.235	0.276	0.281	0.316	0.150	0.249
$melsub$	0.161	0.222	0.094	0.159	0.242	0.227	0.120	0.196
$mfcc$	0.143	0.194	0.079	0.138	0.116	0.114	-0.008	0.074
plp	0.071	0.126	0.009	0.069	0.084	0.047	-0.051	0.026
Mean:	0.162	0.216	0.104	0.161	0.181	0.176	0.052	0.136

indicating some amount of linear relationship between the IF and GMR estimates. Therefore, it is most likely that the negative r_d values are indicative of differences in offset or scale between the IF and GMR estimates of these features.

Finally, it should be noted that perfectly estimating the IF glottal features should not necessarily be the goal of the proposed estimation method, as the IF features themselves can be very noisy. In fact, comparison between Table 14 and Table 4 reveals some agreement between the measurement reliability of the IF features and how closely the GMR method is able to estimate them from the spectral envelope of speech.

Table 11: Mean correlation coefficient r_c between IF and GMR glottal features, by number of GMM components (N_r).

	Males				Females			
N_r	/iy/	/ae/	/ux/	Mean:	/iy/	/ae/	/ux/	Mean:
2	0.375	0.444	0.343	0.387	0.390	0.417	0.328	0.378
4	0.397	0.456	0.355	0.403	0.423	0.422	0.336	0.394
8	0.408	0.463	0.342	0.404	0.436	0.429	0.325	0.397
16	0.404	0.459	0.321	0.395	0.424	0.418	0.309	0.384
Mean:	0.396	0.455	0.340	0.397	0.418	0.421	0.324	0.388

Table 12: Mean coefficient of determination r_d between IF and GMR glottal features, by number of GMM components (N_r).

	Males				Females			
N_r	/iy/	/ae/	/ux/	Mean:	/iy/	/ae/	/ux/	Mean:
2	0.147	0.208	0.125	0.160	0.163	0.187	0.092	0.147
4	0.165	0.219	0.128	0.171	0.191	0.185	0.085	0.153
8	0.172	0.223	0.101	0.165	0.198	0.183	0.051	0.144
16	0.163	0.213	0.061	0.146	0.170	0.150	-0.018	0.101
Mean:	0.162	0.216	0.104	0.161	0.181	0.176	0.052	0.136

Table 13: Correlation coefficient r_c between IF and GMR glottal features for $melsub_{41}$ spectral envelope feature vector and 4-component GMMs.

	Males				Females			
Feature:	/iy/	/ae/	/ux/	Mean:	/iy/	/ae/	/ux/	Mean:
OQ_{LF}	0.526	0.639	0.470	0.545	0.559	0.530	0.470	0.520
α_m	0.485	0.507	0.381	0.458	0.175	0.359	0.151	0.228
Qa	0.466	0.510	0.430	0.468	0.398	0.568	0.443	0.470
Rd	**0.736**	**0.747**	0.697	**0.726**	0.699	0.675	0.487	0.620
NAQ	0.631	0.603	0.609	0.614	0.602	0.593	0.481	0.559
ClQ	0.453	0.518	0.448	0.473	0.510	0.482	0.332	0.441
OQa	0.657	0.692	0.622	0.657	0.699	**0.729**	0.670	0.699
SQ_1	0.454	0.518	0.391	0.454	0.304	0.292	0.202	0.266
SQ_2	0.408	0.393	0.339	0.380	0.280	0.261	0.108	0.216
$SQ_{20\text{-}80}$	0.347	0.380	0.214	0.314	0.509	0.465	0.331	0.435
OQ_{80}	0.509	0.675	0.336	0.507	0.571	0.644	0.428	0.547
OQ_{50}	0.413	0.570	0.267	0.417	0.629	0.682	0.478	0.597
OQ_{20}	0.233	0.247	0.213	0.231	0.099	0.120	0.125	0.114
$H1\text{-}H2$	0.632	0.626	0.618	0.625	**0.786**	**0.755**	0.699	**0.747**
HRF	**0.856**	**0.874**	**0.836**	**0.855**	**0.749**	**0.794**	0.664	**0.736**
$TILT$	0.587	0.566	0.514	0.556	**0.733**	**0.774**	0.627	**0.711**
Mean:	0.524	0.567	0.462	0.518	0.519	0.545	0.418	0.494

Table 14: Coefficient of determination r_d between IF and GMR glottal features for $melsub_{41}$ spectral envelope feature vector and 4-component GMMs.

	Males				Females			
Feature:	/iy/	/ae/	/ux/	Mean:	/iy/	/ae/	/ux/	Mean:
OQ_{LF}	0.265	0.399	0.179	0.281	0.294	0.240	0.160	0.231
α_m	0.232	0.245	0.117	0.198	-0.020	0.108	-0.170	-0.027
Qa	0.205	0.249	0.159	0.204	0.095	0.313	0.023	0.144
Rd	**0.540**	**0.546**	0.471	**0.519**	0.487	0.441	0.145	0.358
NAQ	0.391	0.355	0.340	0.362	0.350	0.335	0.118	0.268
ClQ	0.160	0.243	0.163	0.189	0.253	0.216	-0.029	0.147
OQa	0.431	0.478	0.345	0.418	0.482	**0.511**	0.420	0.471
SQ_1	0.200	0.260	0.108	0.189	0.038	-0.004	-0.217	-0.061
SQ_2	0.159	0.119	0.072	0.117	0.014	-0.036	-0.290	-0.104
$SQ_{20\text{-}80}$	0.050	0.003	-0.078	-0.008	0.182	0.099	0.005	0.096
OQ_{80}	0.240	0.444	0.028	0.237	0.314	0.357	0.096	0.256
OQ_{50}	0.148	0.310	-0.019	0.146	0.388	0.450	0.118	0.319
OQ_{20}	0.034	0.033	0.000	0.022	-0.160	-0.160	-0.281	-0.200
$H1\text{-}H2$	0.399	0.387	0.376	0.387	**0.614**	**0.569**	0.479	**0.554**
HRF	**0.732**	**0.763**	**0.696**	**0.730**	**0.558**	**0.629**	0.381	**0.523**
$TILT$	0.343	0.318	0.251	0.304	**0.530**	**0.595**	0.326	0.484
Mean:	0.283	0.322	0.201	0.269	0.276	0.292	0.080	0.216

Table 15: Correlation coefficient r_c and coefficient of determination r_d between RAPT and GMR pitch (f_0) estimates for $melsub_{41}$ spectral envelope feature vector and 4-component GMMs.

	Males				Females			
	/iy/	/ae/	/ux/	Mean:	/iy/	/ae/	/ux/	Mean:
r_c	0.927	0.862	0.886	0.892	0.896	0.826	0.870	0.864
r_d	0.859	0.742	0.785	0.795	0.796	0.677	0.740	0.738

7.2 Pitch and Delta Features

In speech analysis applications that typically use spectral envelope features, it is common practice to augment the feature vector with its first or second-order time derivatives (*delta features*) as a way to capture information about speech dynamics in a form that can be used with static learning models such as GMMs. Given a sequence of N-dimensional SEF observation vectors $\mathbf{x}_k = [x_{k,1}\ x_{k,2}\ ...\ x_{k,N}]^T$, where k is the frame index, an estimate of the first-order derivative for the k^{th} frame can be computed as [125]

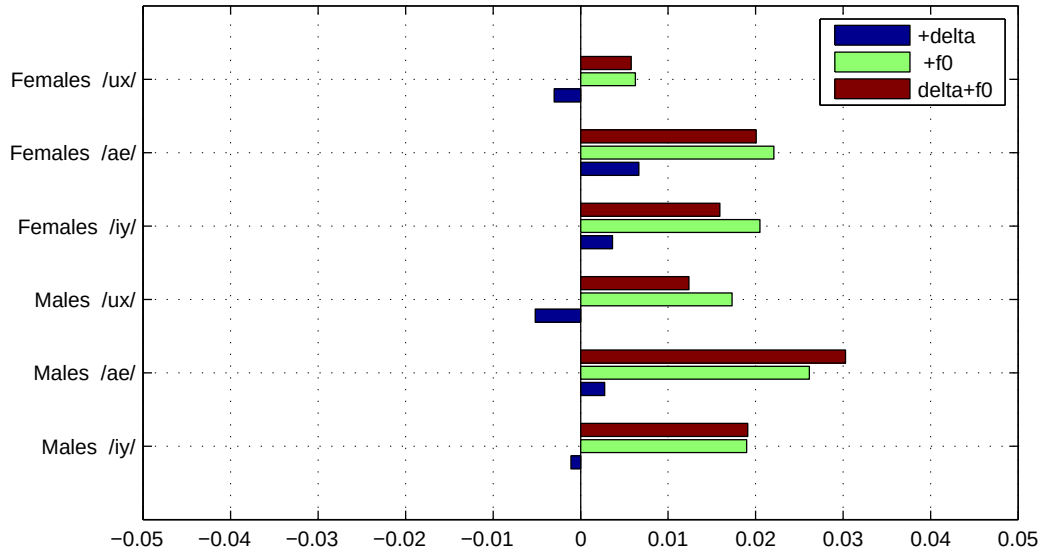
$$\mathbf{d}_k = \frac{1}{10} \sum_{i=1}^{2} i(\mathbf{x}_{k+i} - \mathbf{x}_{k-i}). \tag{43}$$

The derivative vector is then concatenated with the SEF vector to produce the augmented feature set $\phi_k = [\mathbf{x}_k^T\ \mathbf{d}_k^T]^T$. Pitch is another common feature in speech analysis, and it is one of the few properties of the glottal waveform that can be well observed directly from the speech signal. Because pitch is known to be related to changes in other glottal waveform parameters, and since pitch and delta features are commonly used and easily obtainable from the acoustic speech signal, it is reasonable to investigate whether their use in combination with SEFs facilitates glottal feature estimation by the proposed GMR method. To this end, GMMs for the estimation of each glottal feature were trained separately for each gender and phoneme on the TRAIN1 speaker set, using $melsub_{41}$ features either by themselves or augmented by f_0 or a delta feature vector.

Summary results are given in Figure 15 show the average difference in r_c and r_d due to the addition of f_0 and the first-order derivative of the SEF, where r_c and r_d were computed on the TRAIN2 speaker set. These results indicate that the use of f_0 produces a small but consistent improvement in the ability of the GMMs to estimate IF features, while there is no visible pattern for the addition of delta features. Statistical hypothesis tests on $\bar{r}_d$ confirm these observations (two-tailed paired t-tests, DF = 65 for males, DF = 27 for females), with $p < 10^{-30}$ when the baseline SEF

vector is compared to its augmentation by f_0, and no significant difference when the baseline is compared to its augmentation by delta features. Appendix A.2 provides detailed results, showing the values of r_c and r_d for each feature, phoneme, gender, and number of GMM components, as obtained by augmenting the $melsub_4 1$ SEF with f_0 and/or delta features.

These results are not surprising. The inclusion of delta features nearly doubles the dimensionality of the GMM, which increases number of GMM parameters by more than a factor of three, thus imposing additional requirements on the size of the training data set. This explanation is suggested by the results in Figure 15, which show that for the /ux/ phoneme, which is the one with the fewest observations, the addition of delta features actually decreases r_c and r_d. Meanwhile, although pitch is known to have a strong relationship with certain glottal features, it has also been shown that it is possible to estimate pitch from SEFs (Section 7.1.3), so that the addition of explicit pitch information is not expected to be helpful inasmuch as it is already contained in some form within the SEF vector.



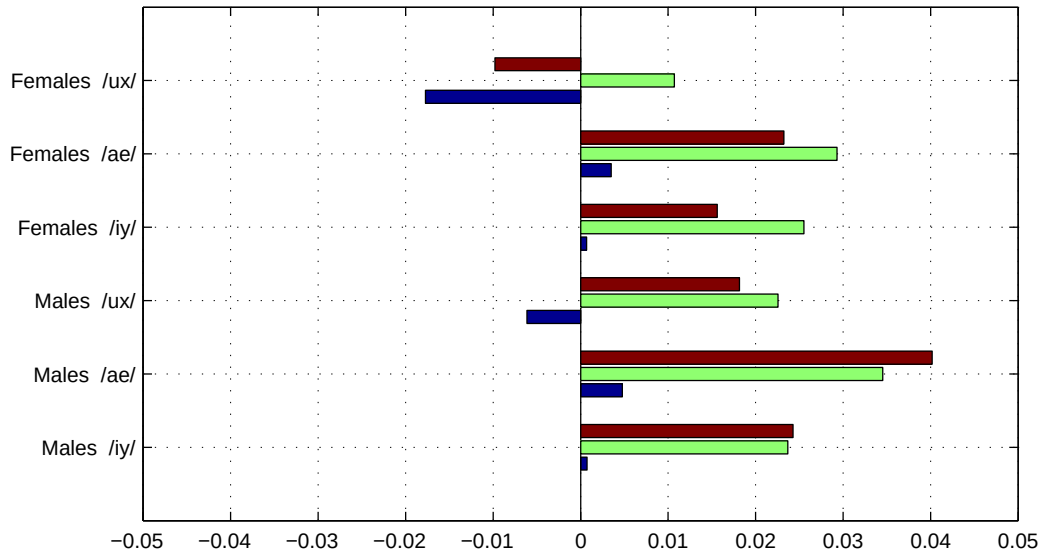


Figure 15: Average difference in Correlation coefficient r_c (a) and Coefficient of determination r_d (b) between IF and GMR glottal features, arising from the addition of delta features and pitch to baseline $melsub_{41}$ feature vector. Positive values indicate increases in r_d or r_c due to the addition of f_0 and/or delta features.

7.3 Measurement Reliability and Statistics of GMR Estimates

Given the overall low measurement reliability of IF glottal features (Section 6.4), an improvement in the quality of the feature estimates may not necessarily be evidenced by an increase in the similarity between IF and GMR features as measured in Section 7.1, since an alternate feature estimation approach that perfectly mimics IF features would only be reproducing errors inherent in the inverse filtering process. Then, it is plausible for GMR features that only roughly correspond to their IF counterparts to be more useful than IF features if this lack of correspondence is accompanied by an increase in measurement consistency. This section presents the measurement reliability of the GMR glottal features as an additional method for assessing the merit of the proposed method of glottal feature estimation.

Based on the results in Section 7.1, the work presented in this section focuses on GMR features obtained by the transformation of $melsub_{41}$ SEF vectors using GMMs with 4 components. Because the data in the TRAIN1 and TRAIN2 speaker sets were used to select the optimal SEF and number of GMM components (Section 7.1), measurement reliability was computed on GMR features extracted from the TEST speaker set, using GMMs that were trained on a combination of the TRAIN1 and TRAIN2 speakers. The purpose of using the independent TEST set was to prevent "parameter tuning" from optimistically biasing the results due to finite observation effects [53]. Just as was done in Section 6.4, the glottal feature observations were segmented into continuous voicings of each phoneme. For each phoneme realization i, the glottal feature observation v_i from the center frame of the segment was paired with the observation u_i from the frame immediately following, and $r_c(u, v)$, $r_d(u, v)$ were computed.

Figures 16 and 17 show the resulting r_c and r_d values, respectively, for GMR features and corresponding IF features obtained using the inverse filtering methods listed

in Table 4. GMR and IF features were computed from the TEST observations. From the figures, it can be observed that the GMR features obtained consistently higher values of r_c and r_d than the IF features. While several IF features obtained values of r_d below 0, all but one GMR feature obtained values of $r_d > 0.5$. Likewise, most of the GMR features obtained a correlation coefficient > 0.8. It is also interesting to note the pattern between the results for the IF and GMR features. For the most part, the least reliable IF features were also the least reliable GMR features, and vice-versa. This behavior is expected, as a noisier IF feature can be expected to result is a less accurate joint PDF estimate, thus impeding the transformation of the SEF vector into the glottal feature via GMR.

To illustrate how GMR features approximate their IF counterparts, a random utterance for each gender was selected and the glottal feature values for each GMR feature were plotted along with their corresponding IF features as a function of time (Figures 18–20). From these figures it can be seen that the statistical approach of the proposed glottal feature estimation procedure results in a type of filtering operation on the feature values, such that the GMR features generally follow the trends of the IF features, but extreme peaks and valleys are filtered out. This is particularly evident for the GMR features that were shown in Tables 13–14 to most closely agree with their corresponding IF features.

The behavior of GMR estimation under adverse conditions can be understood from inspection of the regression formula of Equation 5. The GMR feature estimate $\hat{y}$ is computed from the input feature set $\mathbf{x}$ as the conditional expectation $E(y|\mathbf{x})$, which is in turn estimated by the Gaussian mixture model of the joint PDF between $\mathbf{x}$ and y. Thus, in the extreme case where $\mathbf{x}$ and y are statistically independent, and given a perfect model of the joint density, the optimum estimate of y would simply be $\hat{y} = E(y|\mathbf{x}) = E(y)$, or the mean of y over the GMM training set. Equation 5 further shows that the local (GMM component-specific) mean of y is modified by a

linear term that is a function of $\mathbf{x}$, but this term is weighted by Σ_{iYX}, the covariance matrix between the input and target features. Thus, the regression function suggests that when there is reduced covariance between $\mathbf{x}$ and y, which can arise due to actual independence between the glottal feature and the SEF vector as well as independent noise in the glottal features estimates, the GMR output will elegantly tend towards the mean of y over the training data.

This property of the GMR procedure is further supported by a comparison between the feature statistics of the IF and GMR features. Tables 16 and 17 show the mean and standard deviation of the IF and GMR features on the TEST set for males and females, respectively. These results indicate that all glottal features obtained similar mean values for GMR and IF methods, but the standard deviations for the GMR features were substantially lower. Furthermore, the ratio between the standard deviations of the IF and GMR estimates were generally lower (0.3 – 0.5) for features with low r_d similarity values (Table 14) than those obtained for glottal features where the GMR method was better able to match the IF features (Rd, and HRF for males; $H1$-$H2$, HRF, and $TILT$ for females), which have standard deviation ratios in the ranging from 0.7 to above 0.8. Similarly the standard deviation ratios for the highly reliable and predictable pitch feature were as high as 0.95.

Examination of the pairwise Spearman rank correlations r_r between the GMR features (Tables 18 and 19) indicates an overall increased correlation between features, with preservation of the patterns seen in IF features (discussed in Section 6.4.1) and the addition of some interesting relationships. Rank correlations between feature pairs that are expected to be related to each other were enhanced without exception, these included SQ_1 and SQ_2, OQ_{80} and OQ_{50}, $TILT$ and HRF, as well as NAQ and ClQ. Some pairs of features that were expected to be correlated, but whose IF estimates showed little correlation, obtained much higher values in the GMR case. The LF-model open quotient OQ_{LF} is now correlated with OQ_{80} and OQ_{50}. Similarly,

Rd shows moderate correlation (> 0.60) with Qa as well as NAQ, ClQ and OQa for males, and strong correlation (> 0.80) for females. There is also strong correlation between Rd and OQ features for females. These relationships are expected because Rd was introduced in [40] as a single parameter that is related to the other three wave-shape parameters of the LF model.

The LF-model measure of cycle asymmetry, α_m, now shows correlation to SQ_1 and SQ_2 (-0.62 and -0.64, respectively) for male speakers . The negative correlation is surprising, but could be caused by differences in the definitions of these parameters, as the speed quotients include in their computation of the closing phase the portion of the glottal cycle that is considered the "return phase" in the LF-model. Furthermore, the method of detecting the start of the closed phase t_c as the first zero crossing beyond the GCI t_e (see Section 6.2.2.1), which was used in the computation of SQ_1 and SQ_2, may run into robustness issues due to noise or small DC offsets in the glottal waveform estimates.

It should be noted that unexpected correlations have appeared as well: Qa shows strong correlation to NAQ and ClQ, while OQ_{80} is now strongly correlated with the closing quotient features ClQ, NAQ, and OQa. This last relationship was only observed for female speakers in the IF features. For female speakers, Qa showed strong correlation to all open quotient features except OQ_{20}, and NAQ/ClQ showed strong correlation to OQ_{50} as well as OQ_{80}.

There are at least three possible ways to explain how a set of separate regressors has increased the pairwise correlation between glottal feature estimates. First, there is the possibility that, under a situation where the SEF $\mathbf{x}$ has little useful information about y, the GMR function converges approximately to an estimate of μ_y, but that the small variations that remain in the output $\hat{y}$ mainly reflect changes in $\mathbf{x}$ (Equation 5), giving rise to a general increase in correlations due to the fact that the same sequence of $\mathbf{x}$ observations was used to test each GMR model. This scenario may be occurring

for a few of the glottal features, but is unlikely for most of the features given the positive values of r_d between IF and GMR estimates (Table 14), as well as standard deviations for the GMR features that are only moderately smaller than those obtained for their IF counterparts (Tables 18 and 19).

The second possibility is that the GMR procedure can only estimate the changes in y that are reflected unambiguously in the spectral envelope of the speech signal. In that case, distinct glottal features that have some similarity (e.g. speed quotient and closing quotient) could be forced to converge if their underlying differences are not predictable from the speech spectrum. A third contributing cause for the observed results is that the GMR procedure may actually be improving the IF estimates, since a statistical model that produces estimates based on encapsulated knowledge of a large training dataset should be less likely to produce spurious errors than an inverse filtering procedure that can only make decisions based on a 25 millisecond observation. The removal of spurious errors from the IF estimates could be revealing true underlying relationships among the set of 16 glottal features, which is quite plausible given the low-dimensionality of parametric glottal waveform models (2-3 independent shape parameters) [36, 12, 118], as well as production and perception studies that have shown that a single derived glottal parameter can approximately represent a continuum of pressed to breathy of voice quality [40, 35].

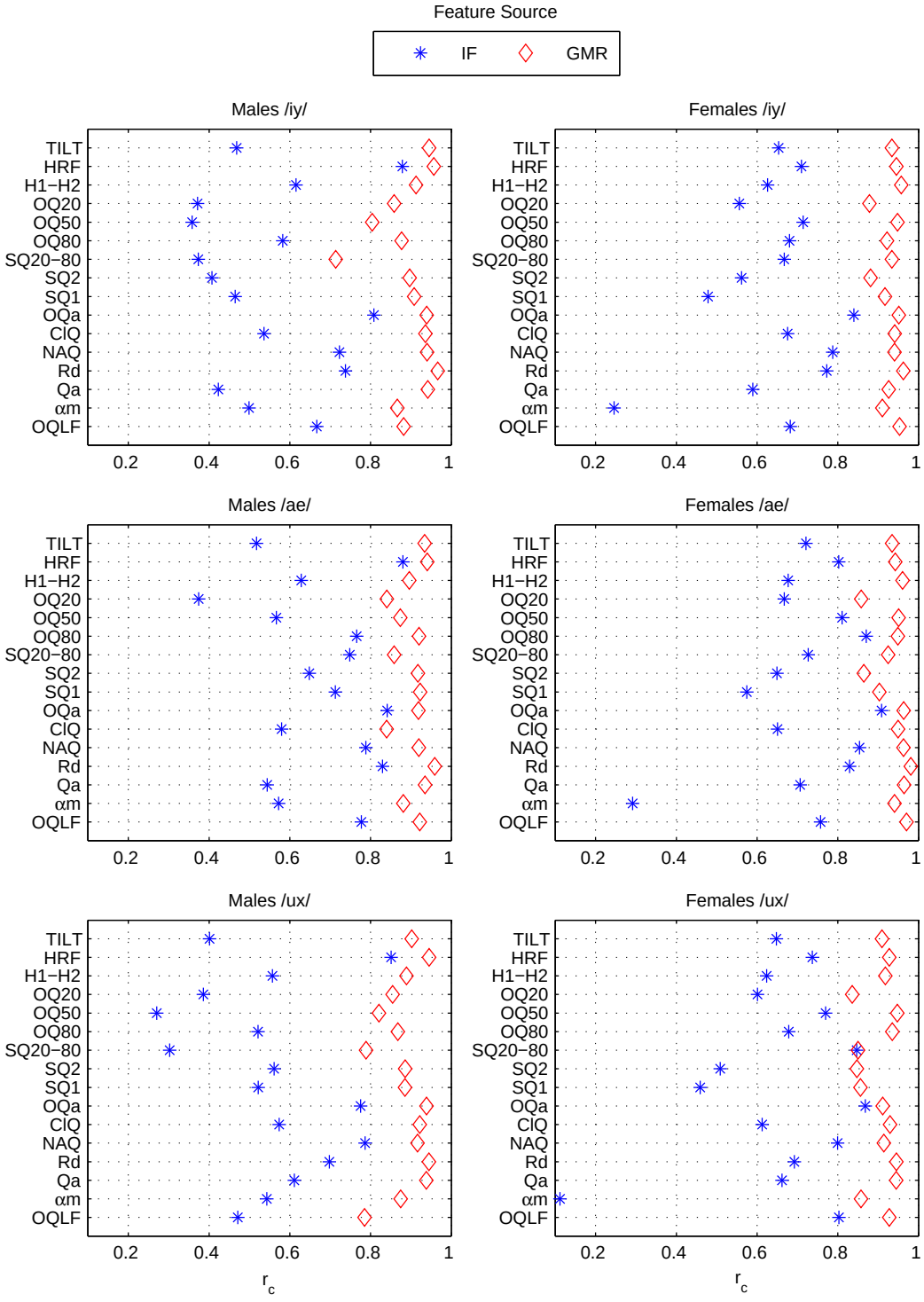


Figure 16: Correlation coefficient (r_c) between observation pairs from adjacent frames. Males and female speakers, TEST dataset, sustained vowels /ae/, /iy/, /ux/. Plot symbols denote feature estimation method: inverse filtering (IF) or GMM regression (GMR).

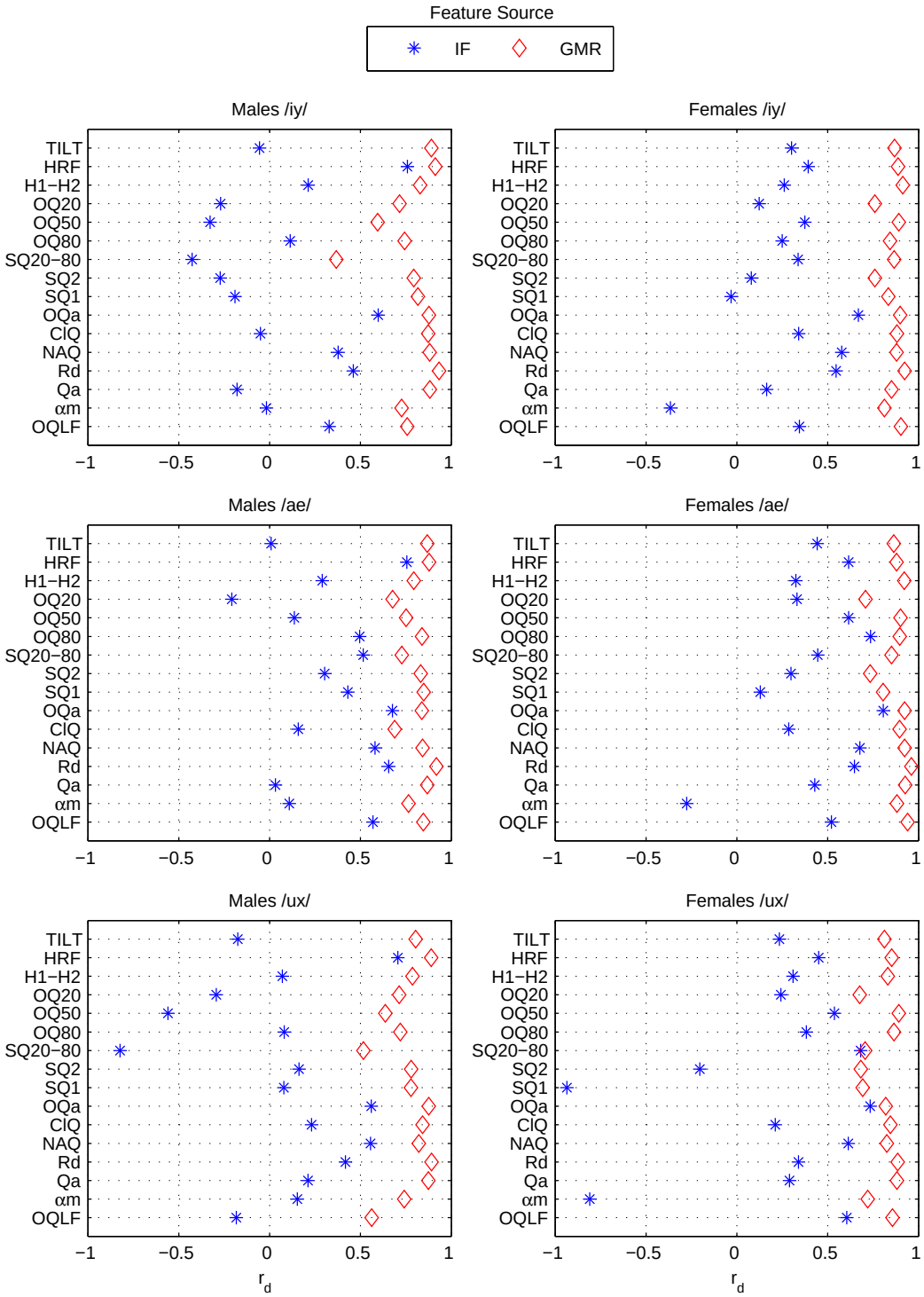
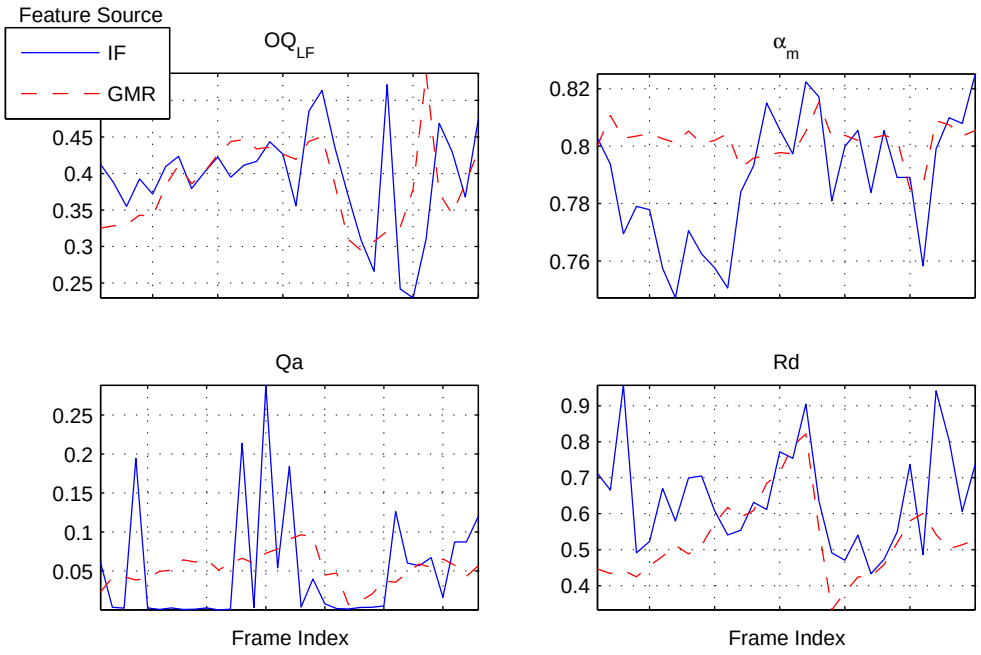
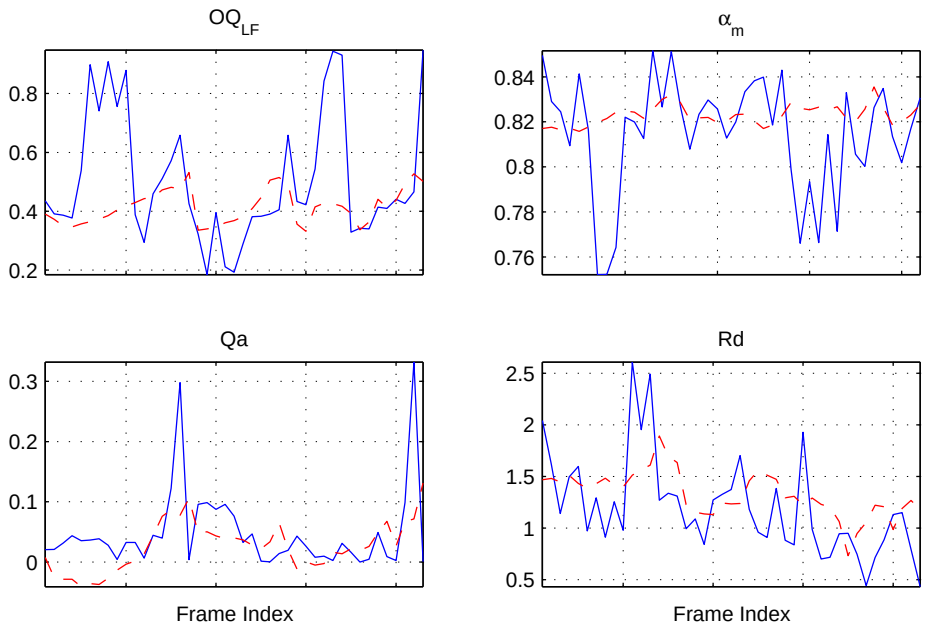


Figure 17: Coefficient of determination (r_d) between observation pairs from adjacent frames. Males and female speakers, TEST dataset, sustained vowels /ae/, /iy/, /ux/. Plot symbols denote feature estimation method: inverse filtering (IF) or GMM regression (GMR).



(a) Male Speaker



(b) Female Speaker

Figure 18: LF-model features for a randomly selected male utterance (a) and female utterance (b) from the TEST dataset. Comparison between features obtained via inverse filtering (IF) and GMM regression (GMR).

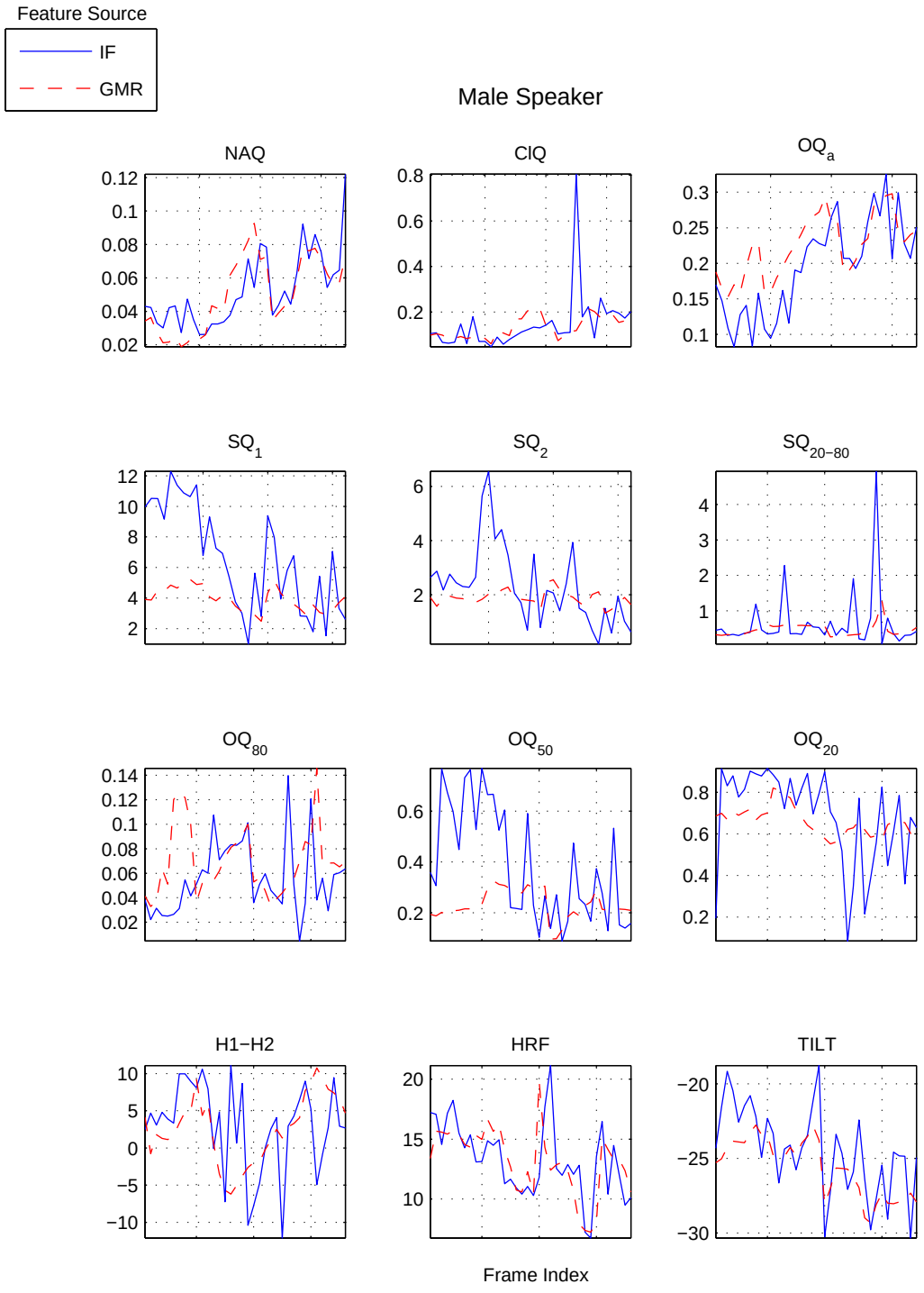


Figure 19: Direct measurement features for a randomly selected male utterance from the TEST dataset. Comparison between features obtained via inverse filtering (IF) and GMM regression (GMR).

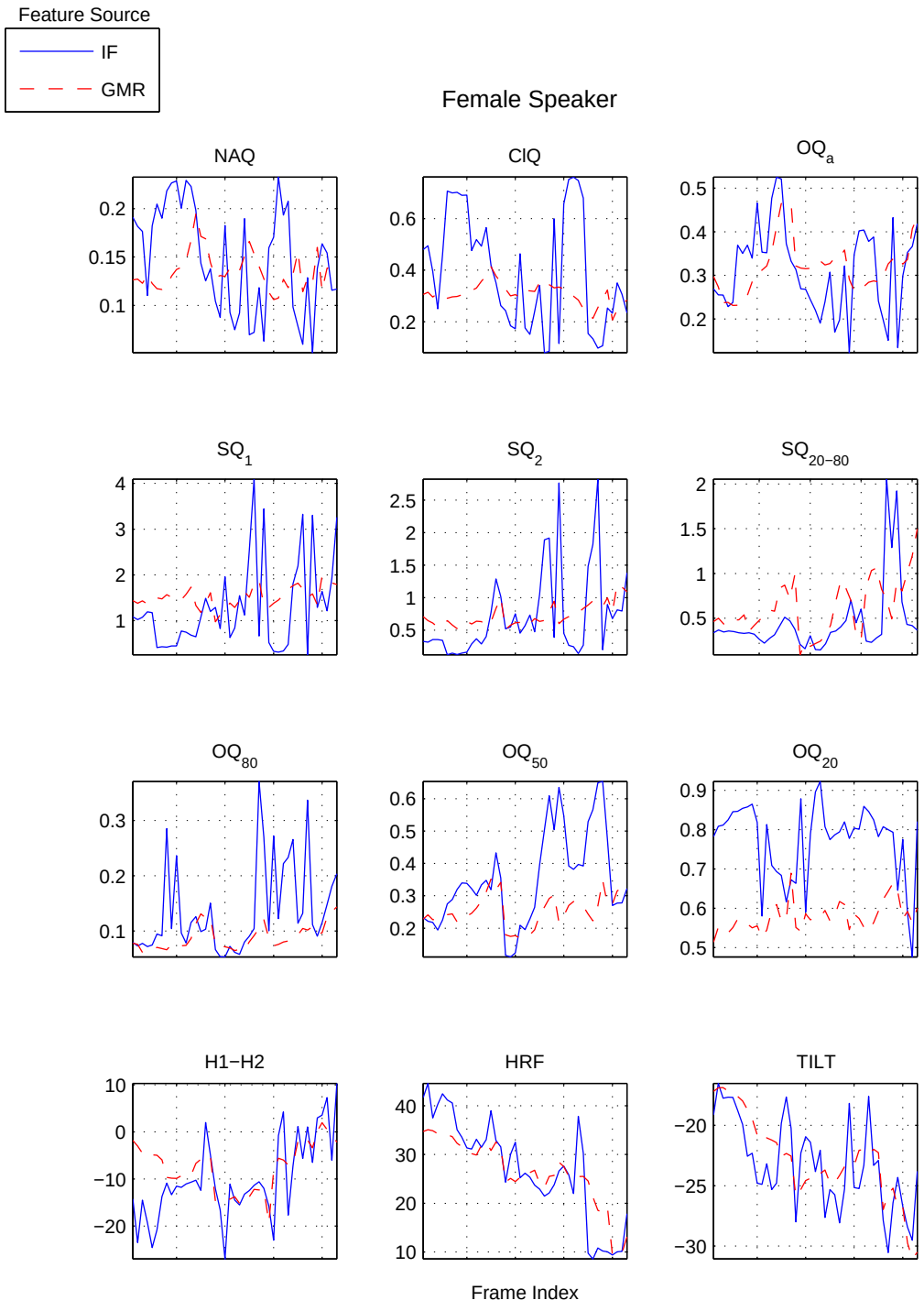


Figure 20: Direct measurement features for a randomly selected female utterance from the TEST dataset. Comparison between features obtained via inverse filtering (IF) and GMM regression (GMR).

Table 16: Mean value and standard deviation (shown in parentheses) for each glottal feature, computed on the TEST dataset from glottal waveforms obtained either by inverse filtering (IF) or by Gaussian mixture regression (GMR) using $melsub_{41}$ spectral envelope feature vector and 4-component GMMs. Values for pitch estimates obtained via the RAPT algorithm or by the GMR method are also given. Male speakers.

Males		/iy/		/ae/		/ux/	
f_0	RAPT	121	(24)	118	(24)	125	(26)
(Hz)	GMR	121	(23)	118	(21)	125	(23)
OQ_{LF}	IF	0.382	(0.11)	0.359	(0.090)	0.379	(0.11)
	GMR	0.385	(0.063)	0.356	(0.062)	0.384	(0.070)
α_m	IF	0.810	(0.021)	0.813	(0.017)	0.809	(0.019)
	GMR	0.809	(0.010)	0.811	(0.0095)	0.809	(0.010)
Qa	IF	0.0408	(0.052)	0.0244	(0.027)	0.0387	(0.049)
	GMR	0.0400	(0.027)	0.0238	(0.016)	0.0373	(0.023)
Rd	IF	0.462	(0.20)	0.419	(0.17)	0.477	(0.19)
	GMR	0.463	(0.14)	0.427	(0.14)	0.484	(0.14)
NAQ	IF	0.0692	(0.029)	0.0588	(0.025)	0.0718	(0.032)
	GMR	0.0683	(0.019)	0.0591	(0.017)	0.0731	(0.023)
ClQ	IF	0.172	(0.087)	0.144	(0.071)	0.186	(0.087)
	GMR	0.171	(0.053)	0.146	(0.043)	0.187	(0.059)
OQa	IF	0.256	(0.067)	0.222	(0.061)	0.276	(0.070)
	GMR	0.256	(0.041)	0.224	(0.041)	0.274	(0.047)
SQ_1	IF	3.22	(1.9)	2.78	(1.3)	3.24	(1.9)
	GMR	3.25	(0.85)	2.77	(0.73)	3.39	(1.1)
SQ_2	IF	1.77	(1.1)	1.43	(0.79)	1.91	(1.2)
	GMR	1.79	(0.49)	1.49	(0.37)	1.93	(0.57)
$SQ_{20\text{-}80}$	IF	0.449	(0.44)	0.370	(0.28)	0.481	(0.53)
	GMR	0.443	(0.24)	0.388	(0.18)	0.476	(0.28)
OQ_{80}	IF	0.0757	(0.044)	0.0678	(0.033)	0.0788	(0.045)
	GMR	0.0748	(0.026)	0.0702	(0.024)	0.0790	(0.025)
OQ_{50}	IF	0.199	(0.12)	0.166	(0.081)	0.205	(0.12)
	GMR	0.198	(0.061)	0.170	(0.053)	0.212	(0.058)
OQ_{20}	IF	0.640	(0.19)	0.661	(0.18)	0.620	(0.18)
	GMR	0.652	(0.055)	0.663	(0.058)	0.635	(0.058)
$H1$-$H2$	IF	1.47	(8.5)	2.82	(7.9)	1.37	(8.6)
(dB)	GMR	1.58	(5.2)	2.96	(5.1)	1.35	(5.6)
HRF	IF	9.60	(6.6)	7.99	(6.1)	10.5	(6.9)
	GMR	9.47	(5.5)	8.00	(5.3)	10.6	(5.8)
$TILT$	IF	-29.4	(4.2)	-29.4	(4.2)	-29.4	(4.1)
(dB/dec)	GMR	-29.3	(2.5)	-29.5	(2.4)	-29.2	(2.2)

Table 17: Mean value and standard deviation (shown in parentheses) for each glottal feature, computed on the TEST dataset from glottal waveforms obtained either by inverse filtering (IF) or by Gaussian mixture regression (GMR) using $melsub_{41}$ spectral envelope feature vector and 4-component GMMs. Values for pitch estimates obtained via the RAPT algorithm or by the GMR method are also given. Female speakers.

Females		/iy/		/ae/		/ux/	
f_0	RAPT	206	(44)	194	(43)	209	(37)
(Hz)	GMR	204	(42)	194	(39)	207	(36)
OQ_{LF}	IF	0.444	(0.13)	0.366	(0.099)	0.489	(0.16)
	GMR	0.449	(0.085)	0.370	(0.066)	0.463	(0.093)
α_m	IF	0.828	(0.017)	0.825	(0.016)	0.825	(0.018)
	GMR	0.827	(0.0054)	0.825	(0.0064)	0.825	(0.0058)
Qa	IF	0.0825	(0.071)	0.0909	(0.073)	0.0735	(0.049)
	GMR	0.0838	(0.039)	0.0913	(0.044)	0.0847	(0.038)
Rd	IF	1.59	(0.66)	1.24	(0.51)	1.66	(0.63)
	GMR	1.58	(0.47)	1.26	(0.37)	1.70	(0.45)
NAQ	IF	0.158	(0.067)	0.121	(0.052)	0.180	(0.066)
	GMR	0.155	(0.044)	0.121	(0.033)	0.173	(0.046)
ClQ	IF	0.340	(0.14)	0.269	(0.12)	0.377	(0.13)
	GMR	0.332	(0.079)	0.263	(0.060)	0.356	(0.080)
OQa	IF	0.387	(0.12)	0.315	(0.084)	0.403	(0.11)
	GMR	0.385	(0.086)	0.317	(0.066)	0.405	(0.085)
SQ_1	IF	1.34	(0.75)	1.33	(0.56)	1.23	(0.62)
	GMR	1.36	(0.34)	1.38	(0.25)	1.21	(0.30)
SQ_2	IF	0.706	(0.49)	0.726	(0.39)	0.579	(0.41)
	GMR	0.690	(0.18)	0.743	(0.16)	0.607	(0.20)
$SQ_{20\text{-}80}$	IF	0.682	(0.60)	0.460	(0.38)	0.829	(0.71)
	GMR	0.747	(0.38)	0.509	(0.26)	0.786	(0.36)
OQ_{80}	IF	0.104	(0.054)	0.0768	(0.030)	0.117	(0.063)
	GMR	0.102	(0.031)	0.0803	(0.023)	0.112	(0.038)
OQ_{50}	IF	0.272	(0.11)	0.198	(0.066)	0.283	(0.094)
	GMR	0.263	(0.067)	0.202	(0.048)	0.278	(0.068)
OQ_{20}	IF	0.584	(0.17)	0.554	(0.17)	0.579	(0.18)
	GMR	0.553	(0.060)	0.541	(0.065)	0.542	(0.089)
$H1$-$H2$	IF	-5.48	(8.1)	-5.35	(8.4)	-5.75	(7.3)
(dB)	GMR	-5.58	(6.3)	-5.70	(6.2)	-5.59	(5.4)
HRF	IF	28.5	(7.7)	29.0	(9.1)	29.4	(6.3)
	GMR	28.5	(5.9)	28.5	(7.2)	28.9	(5.2)
$TILT$	IF	-22.5	(4.6)	-22.3	(4.8)	-22.7	(4.0)
(dB/dec)	GMR	-22.4	(3.7)	-22.2	(3.9)	-22.6	(3.0)

Table 18: Mean Spearman rank correlation coefficient r_r between pairs of glottal features obtained via the GMR method using $melsub_{41}$ spectral envelope feature vector and 4-component GMMs. Male speakers.

	f_0	OQ_{LF}	α_m	Qa	Rd	NAQ	ClQ	OQa	SQ_1	SQ_2	SQ_{20-80}	OQ_{80}	OQ_{50}	OQ_{20}	$H1$-$H2$	HRF	$TILT$
f_0	1.00	-.05	.18	.44	**.71**	.45	.45	.35	-.48	-.43	.10	.17	.14	-.38	.15	**.78**	**.73**
OQ_{LF}	-.05	1.00	-.48	.30	.40	.42	.37	.59	.15	.22	**.75**	.61	.68	-.21	.08	-.35	-.13
α_m	.18	-.48	1.00	.29	-.01	.22	.18	-.07	-.62	-.65	-.54	-.11	-.34	.15	-.06	.29	-.08
Qa	.44	.30	.29	1.00	.64	**.76**	**.71**	.56	-.53	-.47	.26	.61	.48	-.06	.09	.29	.13
Rd	**.71**	.40	-.01	.64	1.00	.68	.67	.64	-.44	-.31	.41	.59	.55	-.26	.23	.42	.48
NAQ	.45	.42	.22	**.76**	.68	1.00	**.86**	**.77**	-.59	-.43	.33	.69	.51	-.18	.13	.16	.04
ClQ	.45	.37	.18	**.71**	.67	**.86**	1.00	**.75**	-.58	-.41	.36	**.70**	.49	-.14	.10	.15	.06
OQa	.35	.59	-.07	.56	.64	**.77**	**.75**	1.00	-.37	-.15	.48	**.72**	.58	-.27	.14	-.05	.00
SQ_1	-.48	.15	-.62	-.53	-.44	-.59	-.58	-.37	1.00	**.79**	.17	-.32	.03	-.01	-.11	-.26	-.12
SQ_2	-.43	.22	-.65	-.47	-.31	-.43	-.41	-.15	**.79**	1.00	.20	-.11	.18	.04	-.03	-.34	-.14
$SQ_{20\text{-}80}$	.10	**.75**	-.54	.26	.41	.33	.36	.48	.17	.20	1.00	.49	.59	-.29	.10	-.17	.08
OQ_{80}	.17	.61	-.11	.61	.59	.69	**.70**	**.72**	-.32	-.11	.49	1.00	**.73**	.05	.13	-.17	-.10
OQ_{50}	.14	.68	-.34	.48	.55	.51	.49	.58	.03	.18	.59	**.73**	1.00	-.05	.07	-.06	.05
OQ_{20}	-.38	-.21	.15	-.06	-.26	-.18	-.14	-.27	-.01	.04	-.29	.05	-.05	1.00	-.06	-.27	-.35
$H1$-$H2$	.15	.08	-.06	.09	.23	.13	.10	.14	-.11	-.03	.10	.13	.07	-.06	1.00	-.08	.03
HRF	**.78**	-.35	.29	.29	.42	.16	.15	-.05	-.26	-.34	-.17	-.17	-.06	-.27	-.08	1.00	**.72**
$TILT$	**.73**	-.13	-.08	.13	.48	.04	.06	.00	-.12	-.14	.08	-.10	.05	-.35	.03	**.72**	1.00

Table 19: Mean Spearman rank correlation coefficient r_r between pairs of glottal features obtained via the GMR method using $melsub_{41}$ spectral envelope feature vector and 4-component GMMs. Female speakers.

	f_0	OQ_{LF}	α_m	Qa	Rd	NAQ	ClQ	OQa	SQ_1	SQ_2	SQ_{20-80}	OQ_{80}	OQ_{50}	OQ_{20}	$H1$-$H2$	HRF	$TILT$
f_0	1.00	.57	.29	.54	**.83**	**.71**	.69	.62	-.46	-.49	.43	.52	.64	.18	.42	**.72**	**.75**
OQ_{LF}	.57	1.00	.45	**.70**	**.71**	**.74**	.67	**.86**	-.15	-.23	**.83**	**.80**	**.83**	.24	.23	.26	.24
α_m	.29	.45	1.00	.66	.35	.37	.31	.51	-.06	.00	.40	.50	.44	.06	.05	.17	.11
Qa	.54	**.70**	.66	1.00	.62	.68	.60	**.78**	-.28	-.23	.65	**.76**	**.70**	.09	.18	.29	.25
Rd	**.83**	**.71**	.35	.62	1.00	**.89**	**.89**	**.82**	-.58	-.53	.54	.69	**.81**	.11	.33	.54	.52
NAQ	**.71**	**.74**	.37	.68	**.89**	1.00	**.93**	**.89**	-.61	-.55	.56	**.78**	**.84**	.10	.22	.43	.37
ClQ	.69	.67	.31	.60	**.89**	**.93**	1.00	**.83**	-.67	-.60	.49	**.73**	**.81**	.09	.20	.42	.37
OQa	.62	**.86**	.51	**.78**	**.82**	**.89**	**.83**	1.00	-.40	-.34	**.72**	**.84**	**.88**	.14	.21	.33	.26
SQ_1	-.46	-.15	-.06	-.28	-.58	-.61	-.67	-.40	1.00	**.74**	-.01	-.32	-.34	.14	-.01	-.36	-.30
SQ_2	-.49	-.23	.00	-.23	-.53	-.55	-.60	-.34	**.74**	1.00	-.10	-.24	-.27	.13	-.04	-.40	-.36
$SQ_{20\text{-}80}$	.43	**.83**	.40	.65	.54	.56	.49	**.72**	-.01	-.10	1.00	**.70**	.69	.12	.21	.19	.18
OQ_{80}	.52	**.80**	.50	**.76**	.69	**.78**	**.73**	**.84**	-.32	-.24	**.70**	1.00	**.87**	.20	.21	.22	.16
OQ_{50}	.64	**.83**	.44	**.70**	**.81**	**.84**	**.81**	**.88**	-.34	-.27	.69	**.87**	1.00	.28	.26	.31	.29
OQ_{20}	.18	.24	.06	.09	.11	.10	.09	.14	.14	.13	.12	.20	.28	1.00	.33	.02	.10
$H1$-$H2$	.42	.23	.05	.18	.33	.22	.20	.21	-.01	-.04	.21	.21	.26	.33	1.00	.19	.38
HRF	**.72**	.26	.17	.29	.54	.43	.42	.33	-.36	-.40	.19	.22	.31	.02	.19	1.00	**.87**
$TILT$	**.75**	.24	.11	.25	.52	.37	.37	.26	-.30	-.36	.18	.16	.29	.10	.38	**.87**	1.00

7.4 Joint Models for Estimation of Glottal Feature Vectors on Multiple Phonemes

Up to this point, the proposed glottal feature estimation procedure has been evaluated separately on single glottal features in an effort to discover feature-specific differences in the merit and accuracy of the estimates while maintaining a low-dimensional model with a correspondingly low number of training parameters. Models were also trained separately on each phoneme as a way to limit the variability in the input SEFs, with the assumption that this would result in a simpler joint density function whose primary source of variation is the interaction between the SEFs and the glottal feature rather than the variations in SEFs across phonemes. Furthermore, the analysis has been limited to three stationary vowels {/iy/, /ae/, /ux/} that are representative of the class of phonemes most amenable to inverse filtering, as discussed in Section 6.3.

However, there is strong motivation for the use of larger models which could simultaneously output several glottal features from any voiced phoneme. Not only would such a system do away with the requirement of a preliminary speech recognition / phonetic alignment step, it may also model more efficiently the variability in the speech spectral envelope and the glottal features for the following reasons: First, there is considerable similarity and coupling between some of the glottal features under study (Section 6.2, Section 6.4.1), which a joint GMM that can represent both local and global covariance could obviously exploit. Second, studies on the spectral variations produced by changes in glottal features [108, 12, 36] have shown certain spectral effects to be the result of the *combined* variation of two or more spectral features, such that the features can be better ascertained by a model that can represent this joint relationship. Finally, it is well known that even if the phonetic content is controlled by processing frames with the same phonetic label, there will still be considerable vocal tract variation in the observed spectra due to factors that

include physiological differences among speakers, changes in pronunciation, and co-articulation [93, 57]. Therefore, training a model with speech from several phonemes may not imply adding much more complexity to a statistical model that already has to cope with these additional sources on spectral variation.

From the description in Section 4.3, it should be clear that the proposed GMM regression method is not limited to estimating a single feature, but can just as easily estimate one feature vector from another. If $\mathbf{x}_m$ and $\mathbf{y}_m$ represent the SEF vector and a vector containing several glottal features for the m^{th} frame, respectively, a GMM modeling their joint distribution can be trained on $\mathbf{w}_m = [\mathbf{x}_m^T \mathbf{y}_m^T]^T$ and used to transform an independent set of SEF vectors $[\hat{\mathbf{x}}_1\, \hat{\mathbf{x}}_2 \ldots \hat{\mathbf{x}}_K]$ into the corresponding glottal feature vector estimates $[\hat{\mathbf{y}}_1\, \hat{\mathbf{y}}_2 \ldots \hat{\mathbf{y}}_K]$, according to Equation 5.

The correlation coefficient r_c and coefficient of determination r_d between GMR and IF features from the TRAIN2 speaker set were computed for each feature and gender, using joint model and separate model configurations. All GMMs were trained on the TRAIN1 speaker set. The r_c and r_d results for a single GMM trained on a vector of 16 glottal features from observations of all three vowels {/iy/, /ae/, /ux/} were compared against the average r_c and r_d values for a 3-GMM configuration (one GMM per phoneme) and a 48-GMM configuration (one GMM per glottal feature per phoneme). The results show that, especially for female speakers, and particularly for features that already attained high values in the baseline separate-GMM configuration, the joint model tends to improve the approximation of GMM-estimated features to their IF counterparts. Detailed results for each glottal feature and each value of N_r are given in Appendix A.3.

To evaluate performance on additional phonemes, GMMs were again trained on the TRAIN1 speaker set either on all stationary phonemes in the TIMIT phonetic code [46] {/iy/, /ih/, /eh/, /ae/, /aa/, /ah/, /ao/, /uh/, /uw/, /ux/, /er/} or on all voiced frames, regardless of phonetic transcription. The set of evaluated model sizes

was expanded to include 32-component and 64-component GMMs in order to take advantage of the increased number of observations due to the additional processed phonemes and to cope with the wider variety of speech spectra. The results for each case are given in Appendix A.4 for the TRAIN2 set, and show substantial improvement due to the use of all phonemes for certain features (e.g. Qa on female speakers, OQ_{80}) and small decreases on others. It is also noticeable from these figures that in the case of additional phonemes, there is a small but consistent advantage to using the larger 32 and 64-component GMMs.

These results are summarized in Table 20, which shows the average r_d values for each model configuration (joint or separate, 3 vowels or additional phonemes) along with the results of statistical significance tests comparing against the r_d values for a baseline separate-GMM configuration. The optimal value of N_r (number of GMM components) was independently chosen for each combination of feature and method. The results show small improvements in r_d when a single GMM was used for all features and all 3 vowels, as well as for the case where additional phonemes were processed. However, the differences are not statistically significant once the confidence level is adjusted for multiple comparisons. Still, the small improvements in $\bar{r}_d$ for the larger GMMs indicate that a single GMM that outputs the entire feature vector for all the processed phonemes may be used without sacrificing the quality of the estimation, and that, optionally, all voiced phonemes may be processed, thus allowing the proposed procedure to be used without the need for a phonetic transcription.

7.5 Speaker Discrimination Ability of IF and GMR Glottal Waveform Features

A third way to compare the merits of a set of glottal feature estimates is to evaluate their performance in a speech analysis application. This evaluation method can be used to determine the practical implications of differences in measurement reliability and in the level of agreement between GMR and IF features. Because glottal waveform

Table 20: Mean r_d for separate and joint GMMs. p-values obtained from paired, two-tailed t-tests (DF = 65 for males, 27 for females) against the baseline condition in the first row (separate GMM per feature-phoneme pair). The optimal value of N_r (number of GMM components) was used for each glottal feature / model type combination.

Model Type:	Males		Females	
	$\bar{r}_d$	p	$\bar{r}_d$	p
48 GMMs: one per feature / phoneme pair	0.296		0.278	
3 GMMs: one per phoneme (/iy/, /ae/, /ux/)	0.280	0.1688	0.268	0.4798
1 GMM: {/iy/, /ae/, /ux/}	0.307	0.2917	0.313	0.0742
1 GMM: all stationary vowels	0.297	0.8992	0.329	0.0111
1 GMM: all voiced phonemes	0.310	0.4631	0.346	0.0031

variation has been shown to contribute to the identification and transformation of voice identity [91, 68, 22, 67], and because the TIMIT corpus contains speech samples from a large number of speakers, this section focuses on a speaker identification task. The goal of these experiments was not to improve upon the performance of existing speaker identification systems, but to measure the ability of each feature or feature vector to distinguish the voices of different speakers. With this in mind, the evaluation was designed around binary classification of speaker pairs.

7.5.1 Procedure

Speaker identification proceeded as follows: For each gender, the 20 speakers with the most observations of the vowels {/iy/, /ae/, /ux/} were selected from the TEST speaker set. There were at least 200 total observations per speaker (from the three vowels combined). Spectral envelope ($melsub_{41}$) and glottal waveform features were computed for each observation. In the case of inverse filtering features, the IF method was selected on a per-feature basis, as shown in Table 4. For GMR features, the regression was performed using 4-component GMMs which had been previously trained on $melsub_{41}$ and IF glottal features from the TRAIN1 and TRAIN2 speaker sets combined.

For a given feature vector ψ, which may contain a single feature h (in which case $\psi = h$) or any combination of SEF or glottal features ($\psi = [h_1\, h_2 \ldots h_N]^T$), the 10 sentences from each speaker $\omega_l,\ l \in \{1 \ldots 20\}$ were divided randomly into three sets of sentences i_1, i_2, and i_3, with an approximately equal number of sentences per set. Pairs of sentence sets were combined into three *folds*: $j_1 = \{i_2, i_3\},\ j_2 = \{i_1, i_3\},\ j_3 = \{i_1, i_2\}$, and used to train the GMMs $f_{\omega j_1}$, $f_{\omega j_2}$, and $f_{\omega j_3}$. The remaining sentence set for each fold was reserved for testing. This data division procedure was repeated to obtain 20 random partitions τ_t of the data. All GMMs were trained using diagonal covariance matrices, as this enables the use of a larger number of GMM components[1] N_c and has been generally found to produce better results than full-covariance matrices in GMM-based speaker identification (for example, see [91, 97, 96]).

For each pair of speakers ω_l and ω_m, and each random partition τ_t, the classification accuracy was computed via 3-fold cross-validation. For each fold n, the observations from the test sentence sets i_n of speakers ω_l and ω_m were concatenated. To ensure a 50% baseline classification rate, a random subset of K_n observations for the most populous speaker was selected in order to match the K_n observations for the least populous speaker. Given the labeled test observations $[\psi_1\ \psi_2\ \ldots\ \psi_{K_n}]$ with speaker labels $[c_1\ c_2\ \ldots\ c_{K_n}]$ and the GMMs for each speaker and each fold $f_{\omega_l j_n}$, $f_{\omega_m j_n}$, the class label estimates $[\hat{c}_1\ \hat{c}_2\ \ldots\ \hat{c}_{K_n}]$ were generated according to Equation 12, and the classification rate $C_{\tau_t w_l w_m j_n}$ was computed according to Equation 13. The mean classification rate $\bar{C}$ for the feature vector ψ was obtained by averaging over the random partitions, folds, and speaker pairs:

$$\bar{C} = \frac{1}{20}\sum_{t=1}^{20}\frac{1}{3}\sum_{n=1}^{3}\frac{1}{190}\sum_{l=1}^{19}\sum_{m=l+1}^{20} C_{\tau_t w_l w_m j_n} \tag{44}$$

[1] $N_c < K/9$, where K is the number of training observations, guarantees a 3:1 ratio between training data elements and GMM parameters when diagonal covariance matrices are used.

7.5.2 Results

Figures 21 and 22 show the mean classification rate $\bar{C}$ of each glottal feature for male and female speakers, respectively. All glottal features obtained classification rates above 60%, thus showing some ability to discriminate between speakers. The figures show that for most glottal features, the GMR method resulted in higher classification rate than inverse filtering. This relationship can be observed more clearly on Figures 23 and 24, which show the difference $\bar{C}_{IF} - \bar{C}_{GMR}$ between the mean classification rates for IF and GMR features to be as high as 7%. The results of statistical significance tests (two-tailed paired t-tests, DF=9) given in Table 21 show that, for female speakers the mean $\bar{C}$ across all the glottal features is significantly higher for GMR features than for IF features. For males, the classification rate was also higher, but the difference was not statistically significant. The results indicate that, at least in the context of single features, GMR features at least equally able to discriminate between speaker voices.

7.5.3 Speaker Discrimination Ability of Glottal Feature Combinations

To examine the discrimination ability of glottal features in a more realistic scenario, the classification rate for a *combination* of glottal features was evaluated by constructing, for each gender, a glottal feature subset chosen from the studied set of 16 glottal features in an effort to represent the main aspects of variation in glottal behavior. The criteria for constructing this feature vector were threefold: First, it should ideally contain at least one representative feature for each of these salient glottal source features: open quotient, speech quotient, closing quotient, and spectral tilt. Second, the chosen features should have the highest possible measurement reliability. Third, correlation among pairs of features should be minimized in order to diminish information redundancy. These criteria were applied by first selecting the highest-reliability representative (Table 4) from each feature type, then pruning the resulting feature set

by marking pairs of features with rank-correlation above 0.70 (Tables 18 and 19) and removing the element with the lower measurement reliability from each pair. $H1$-$H2$ was added to the feature sets, as it does not appear to be strongly correlated with any other feature. The resulting feature sets were $\{OQ_{LF}$, NAQ, Rd, $H1$-$H2$, HRF, $\alpha_m\}$ for males and $\{OQa$, HRF, $SQ2$, $H1$-$H2\}$ for females. Principal component analysis (PCA) was used to obtain uncorrelated coordinates from these feature sets.

In addition to using a combination of glottal features, the evaluation dataset was expanded to include all voiced phonemes, and a single joint GMM was used for each gender to estimate every glottal feature on the entire TEST dataset. The number of speakers was increased to 50, which resulted in at least 1000 observations per speaker. The data division and testing procedure remained identical to that described in Section 7.5.1, namely 20 random repetitions of 3-fold cross-validation.

The mean classification rate for this feature vector is given in Table 22, which shows classification rates as high as 90.3% (males) and 87.7% (females). A very slight increase in classification rate ($\approx 0.3\%$) was observed for the GMR features over IF features for the males, and a 2.4% decrease was observed for the females, but neither difference was statistically significant (paired t-tests, DF=49). This result, in combination with the results for single glottal features (Table 21) further suggests that the GMR features can be as useful for discriminating speaker's voices as the IF features.

The same glottal feature vectors were then used in combination with the $melsub_{41}$ SEF to examine how they complement the speaker discrimination ability of the spectral magnitude envelope. The results, listed in Table 23, show a small, but statistically significant ($p < 1.1 \times 10^{-4}$) increase in classification rate due to the addition of IF glottal features. The small value of the increase (0.4% – 0.9 %) is not surprising. The fact that the GMR procedure was able to transform the SEF vector into

Table 21: Mean pairwise speaker classification rate for individual glottal features, with p values of two-tailed paired t-tests (DF=9), $N_c = 8$.

Feature Type:	Males		Females	
	$\bar{C}$	p	$\bar{C}$	p
IF	0.6469		0.6467	
GMR	0.6616	3.020E-02	0.6693	9.853E-04

glottal features that are equal or better speaker discriminators than their IF counterparts when used alone (Figures 23 and 24) indicates that most speaker-specific glottal source information (with respect to the set of glottal waveform features under study) is already contained in the spectral envelope of the acoustic speech signal. Therefore, one would expect very little discrimination power due to the addition of IF features, and, ideally, no advantage from the addition of GMR features, since the latter were themselves estimated from the SEFs. Indeed, hypothesis tests show no statistically significant improvement over $melsub_{41}$ features due to the addition of GMR features. The small decrease in classification rate of the $melsub_{41}$ + GMR set with respect to the combination $melsub_{41}$ + IF (0.2% for males, 0.7% for females) was found to be statistically significant only for female speakers ($p < 3.5 \times 10^{-5}$).

Finally, it is worth noting that the results for the $melsub_{41}$ feature set were not greatly higher than those obtained using a glottal feature vector by itself. The $melsub_{41}$ features obtained classification rates of 95.2% for males and 92.0% for females, representing an improvement over glottal features of 4.9% and 4.3%, respectively. This modest difference, taken in the context of the limitations and difficulty associated with inverse filtering, highlights the importance of the glottal source as a component of voice identity.

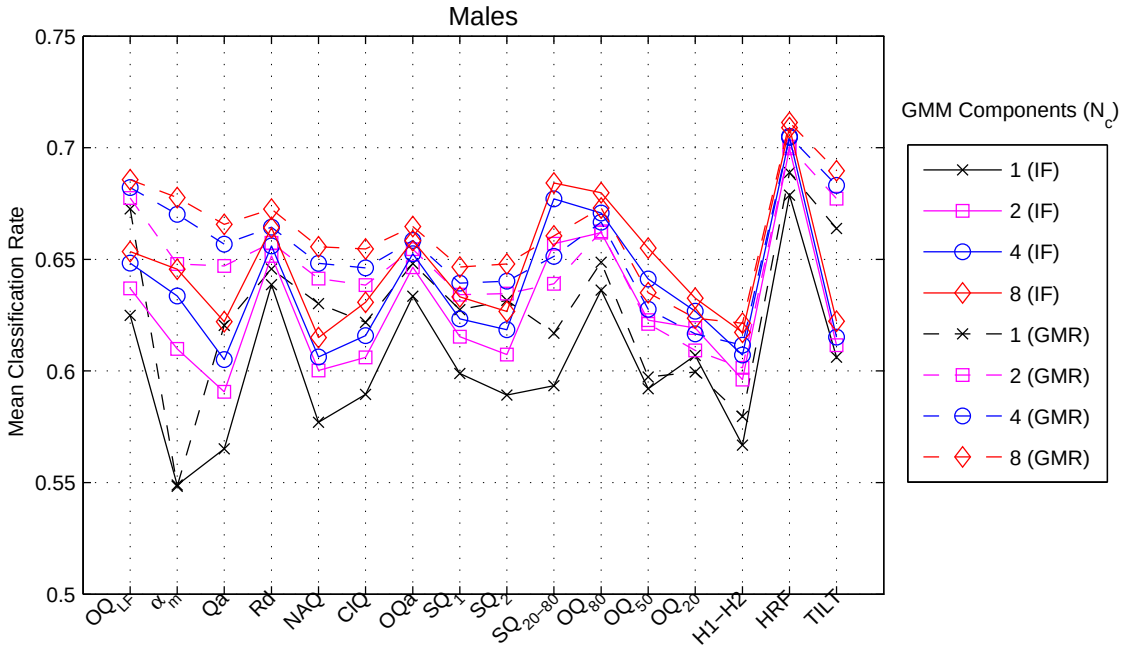


Figure 21: Mean pairwise speaker classification rate for individual inverse filtering (IF) and Gaussian mixture regression (GMR) features, male speakers.

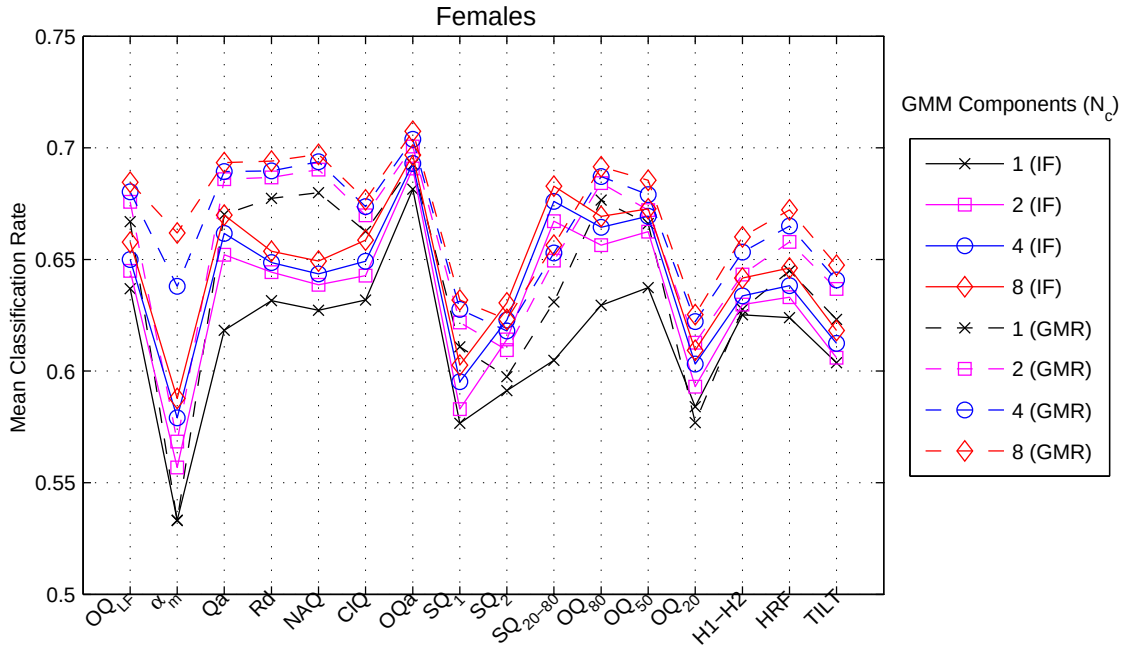


Figure 22: Mean pairwise speaker classification rate for individual inverse filtering (IF) and Gaussian mixture regression (GMR) features, female speakers.

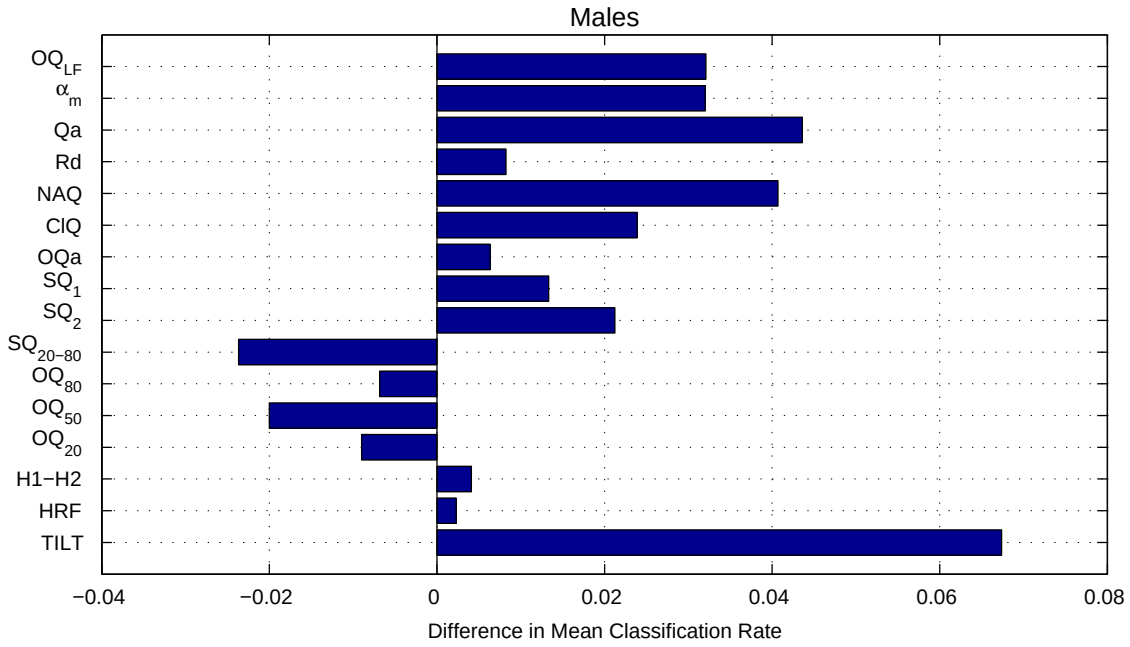


Figure 23: Difference between mean pairwise speaker classification rate for Gaussian mixture regression and inverse filtering features, $N_c = 8$, male speakers. Positive values indicate higher classification accuracy of GMR features with respect to IF features.

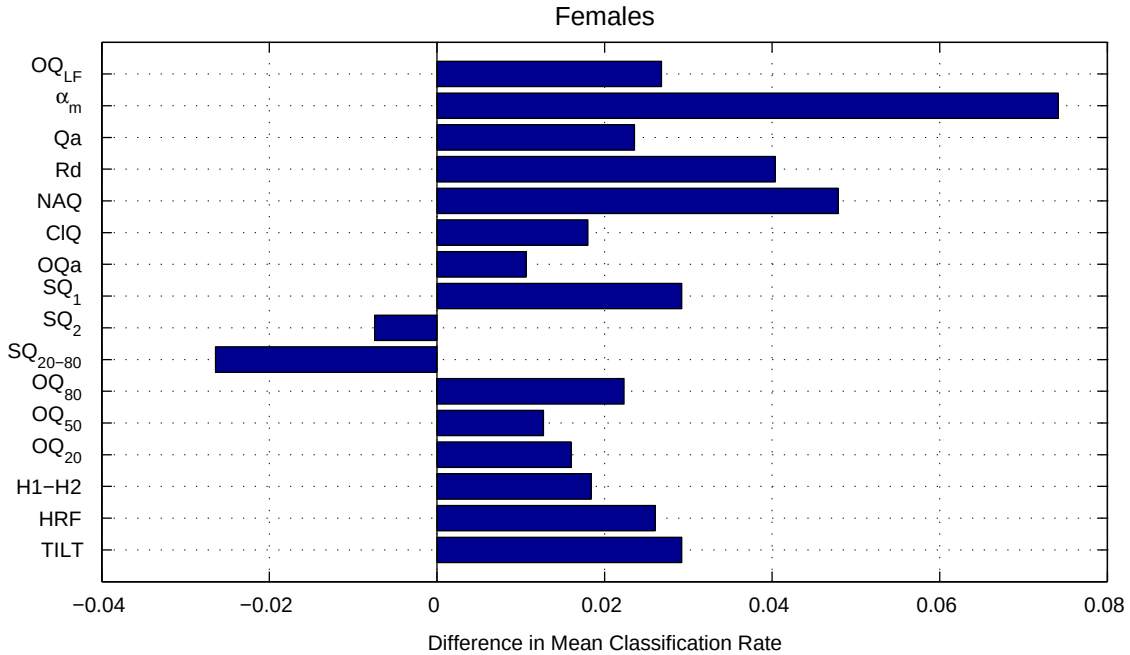


Figure 24: Difference between mean pairwise speaker classification rate for Gaussian mixture regression and inverse filtering features, $N_c = 8$, female speakers. Positive values indicate higher classification accuracy of GMR features with respect to IF features.

Table 22: Mean pairwise speaker classification rate, using a vector of glottal features.

	Feature Source:	GMM Components (N_c)			
		8	16	32	64
Males	IF	0.844	0.864	0.884	0.900
	GMR	0.847	0.867	0.886	0.903
Females	IF	0.811	0.831	0.855	0.877
	GMR	0.776	0.799	0.826	0.853

Table 23: Mean pairwise speaker classification rate, using a combination of spectral envelope features with a glottal feature vector.

	Feature Set	GMM Components (N_c)			
		8	16	32	64
Males	$melsub_{41}$	0.944	0.952	0.955	0.952
	$melsub_{41} + IF$	0.950	0.957	0.960	0.956
	$melsub_{41} + GMR$	0.944	0.953	0.957	0.954
Females	$melsub_{41}$	0.868	0.891	0.911	0.920
	$melsub_{41} + IF$	0.888	0.908	0.924	0.929
	$melsub_{41} + GMR$	0.874	0.895	0.914	0.922

CHAPTER VIII

CONCLUSION

8.1 Research Summary

The objective of the research presented in this thesis was to investigate which characteristics of the glottal source in non-pathological speech are contained in a predictable manner within a spectral envelope representation of the acoustic speech signal and to what extent can these characteristics be measured directly from commonly used feature vectors that represent the spectral envelope of speech. The motivation for this work arises from the difficulties and uncertainty associated with current methods for extracting glottal information from the acoustic speech signal, along with stated evidence about the usefulness of glottal source features for the discrimination of paralinguistic and extralinguistic content in speech.

The lack of availability of a large speech corpus containing an acoustic signal recorded in synchrony with one or more ancillary signals that represent a more direct observation of glottal and vocal fold behavior, as well as unresolved issues regarding the relationship between these signal modalities and glottal airflow (Appendix B), necessitated the use of a baseline consisting of glottal features computed from glottal waveforms which had been obtained by inverse filtering the acoustic speech signal (IF features). Because inverse filtering is an error-prone procedure, the measurement reliability, or consistency of IF features from a set of four inverse filtering algorithms and two methods of model-fitting were compared for the purpose of determining the most adequate way of obtaining each feature as well as an overall estimate of the noise level of the IF feature estimates. Measurement reliability was evaluated using the correlation coefficient r_c and the coefficient of determination r_d to measure the

similarity between pairs of feature estimates from time-adjacent observations located at the center of each continuous realization of the stationary vowels {/iy/, /ae/, /ux/}. On a dataset of 260 male and 108 female speakers from the TIMIT corpus, it was found that measurement reliability varied widely among IF methods and glottal features, with correlation coefficients in the range (0.14 – 0.92) and coefficients of determination in the range (-0.88 – 0.83), where a coefficient of determination below zero is indicative of a mean squared distance between adjacent observations of the feature that is larger than the variance of the feature over the dataset. For the female speakers, the IAIF inverse filtering method consistently produced the most reliable features, and this method resulted in r_d values above zero for all features. For the males, the most reliable method alternated between closed-phase inverse filtering (CPIF) and IAIF, but six out of 16 features obtained r_d values below zero.

Comparison of the glottal feature means across genders suggested a more abducted glottal configuration for female speakers and the presence of incomplete glottal closure, a result consistent with the literature and with the observed advantage of the IAIF method over closed-phase analysis on female speakers. Analysis of the Spearman rank correlation coefficient to detect the presence of monotonic relationships between IF feature pairs revealed both expected and unexpected trends. The LF-model features did not show much correlation to their direct-measurement counterparts, and the amplitude-based open quotient was actually found to depend mostly on the closing quotient measures.

To establish a statistical model that enables the transformation of spectral envelope features (SEFs) into glottal waveform features, a set of Gaussian mixture models (GMMs) were trained on the aforementioned speaker set to model the joint distribution of four common spectral envelope feature vectors and each of the 16 glottal waveform features under study. The evaluated set of spectral feature vectors consisted of mel-frequency cepstral coefficients ($mfcc$), perceptual linear prediction

coefficients (*plp*), decorrelated channel energies from a 21-band mel-scale filter bank (*melsub*), and a higher-resolution version of *melsub* features using a 41-band filter bank ($melsub_{41}$). Each of the 16 glottal features was obtained using the IF method showing the highest measurement reliability for that feature and gender. GMMs were then used to transform SEFs from an independent set of 66 male and 28 female speakers into glottal feature estimates via Gaussian mixture regression (GMR), and an initial performance evaluation was performed by using r_c and r_d to measure the similarity between the IF and GMR-estimated glottal features.

It was found that, on average, the highest values of r_c and r_d were obtained by using the mel-scale filter-bank energies, with additional improvement for the higher resolution 41-band variation. In addition, out of the evaluated values for the GMM size parameter (number of Gaussian components), it was found that 4-component GMMs resulted in the highest average r_d. Using $melsub_{41}$ SEFs and 4-component GMMs, the correlation coefficient r_c varied on a per-feature basis from 0.21 to 0.86 for males and 0.10 to 0.76 for female speakers, and several features (OQ_{LF}, Rd, NAQ, OQa, OQ_{80}, $H1$-$H2$, HRF, $TILT$) obtained a correlation coefficient above 0.5 consistently across genders. The coefficient of determination varied between -0.01 to 0.73 for males and -0.2 to 0.55 for females.

The effect on GMR regression from the augmentation of the $melsub_{41}$ SEF with delta features (across time) and pitch information was also evaluated. It was found that the addition of f_0 resulted in a small but statistically significant ($p < 1 \times 10^{-30}$) improvement on the r_d values between IF and GMR feature estimates (< 0.04). The differences in r_d due to the addition of delta features were not found to be statistically significant. It was hypothesized that the lack of improvement from using delta features may be due to the size of the training data, as the addition of delta features essentially doubles the dimensionality of the GMM.

In an effort to perform a more comprehensive evaluation of the proposed GMR

glottal feature estimation procedure, the measurement reliability of the GMR features was compared to the measurement reliability of the IF features. The GMMs were trained on a set of 326 male and 136 female speakers (TRAIN dataset), and the reliability of IF and GMR features was computed on an independent set of 112 male and 56 female speakers (TEST dataset). The GMR features were found to obtain consistently higher measurement reliability, as evidenced by higher values of r_c and r_d than for the IF features, with r_c and r_d for the GMR features being consistently higher than 0.8 and 0.5, respectively. A time-plot of the IF and GMR glottal feature estimates reveals that the GMR features tend to follow the general trends of their IF counterparts, but that the statistical nature of the regression procedure results in a filtering of sorts, such that some of the sharp variations of the IF features are not found on the GMR feature plots. This result suggests that the proposed method may actually be able to produce more useful glottal feature estimates by filtering out some of the noise present in the IF features that arises from inverse filtering errors.

In the interest of facilitating the use of the proposed glottal feature estimation procedure, the use of joint models for estimating multiple glottal features from multiple phoneme data using a single GMM was explored. Measurements of r_d between GMR and IF features were used to compare the joint models to the case of individual GMMs for each feature and each phoneme. For both genders, there was a small increase in the mean value of r_d due to the use of a single GMM, and an additional increase when the model was trained and tested on all voiced phonemes. Although the differences with respect to the separate-GMM, 3-phoneme case were not statistically significant, the improvement suggests no real disadvantage to using a single GMM to estimate combinations of glottal features from the desired set of phonemes.

As a final measure of merit for the proposed glottal feature estimation procedure, the performance of IF and GMR features was compared on a speaker identification application. Speaker-specific Gaussian mixture models for single glottal features were

trained on the $/iy/, /ae/, /ux/$ utterances for each speaker and maximum likelihood classification of speaker pairs was performed on 20 male and 20 female speakers from the TEST speaker subset of TIMIT. (The GMMs for generating GMR features had been trained with the independent TRAIN speaker set.), The classification rate was estimated using 20 random repartitions of 3-fold cross-validation, and an equal number of test observations were used for both speakers in a pair to maintain a 50% baseline classification rate.

The most useful single feature was HRF for males and OQa for females, both achieving a classification rate slightly above 70%. For most individual glottal features (12 for males, 14 for females), the mean classification rate for individual GMR features showed a small increase (up to 7.4%) over their IF counterparts. On average, the classification rate for GMR features was 66.8% for males and 67.0% for females. This represented an increase of 2.1% for males and 2.4% for females over IF features, but the difference was not found to be statistically significant.

A larger-scale speaker classification experiment using a single, joint GMM for estimating a vector of glottal features from any voiced phoneme was also performed. The increased number of observations allowed for 50 speakers to be evaluated. The use of a vector of glottal features resulted in a mean classification rate of 90.0% for the males and 87.7% for the females, with an improvement of 0.3% and a decrease of 2.4%, respectively, due to the choice of GMR over IF features. The pairwise speaker classification rate using $melsub_{41}$ feature vectors resulted in a classification rate of 95.2% for males and 92.0% for females, with a small, but statistically significant improvement due to the addition of IF glottal features (0.4% males, 0.9% females). The addition of GMR features to $melsub_{41}$ increased the classification rate by only 0.2%, and the difference was not found to be statistically significant.

8.2 Conclusions

The presented research has yielded insights into the reliability of current glottal inverse filtering algorithms as well as the predictability of glottal waveform features from the magnitude spectral envelope of speech. It was found that the measurement reliability of glottal waveform features obtained via inverse filtering, defined as the similarity between adjacent observations from segments of steady phonation, was generally low and varied widely among IF algorithms and glottal features. The oldest and simplest of the inverse filtering algorithms under consideration (closed-phase analysis) was in many cases the one that resulted in the most reliable feature estimates.

Interestingly, the inseparability of the glottal and vocal tract components of speech that causes inverse filtering to be problematic is also what enables the presented approach to work, as glottal source information is still present in spectral envelope features that in some cases were originally designed to enhance phonetic information while minimizing the effects of other sources of variation in the speech signal [30, 61]. The moderate correlation between glottal feature estimates obtained by IF and GMR indicates that the presented GMR glottal feature extraction procedure can roughly approximate most IF features. Perfect correspondence between IF and GMR features, however, is not always a desirable goal, since the IF features themselves were found to be rather noisy. The much improved measurement reliability of GMR features, coupled with their similar performance in a speaker ID application, suggests that the statistical estimation approach of the GMR method incorporates some amount of noise filtering on the IF features, which in some cases resulted in individual GMR features that appeared to be more useful for speaker discrimination than the IF features which they were intended to mimic. It must be noted, however, that the small, but statistically significant advantage of IF features in multivariate experiments suggests that not every aspect of glottal source variation can be captured by the proposed approach.

If it is possible to derive glottal waveform features as a function of the spectral envelope of speech, the question arises as to the usefulness of features obtained via currently available inverse filtering methods in speech analysis applications where the goal is not to characterize glottal source behavior explicitly, but rather to recognize an informational aspect of the speech signal that is believed to be related to glottal source variation. If information characterizing glottal behavior were contained within the speech spectrum in a systematic, predictable manner, then there would be little advantage from the explicit use of glottal features (whether these are obtained via IF or GMR) over the case where the intended analysis is performed directly on the spectral envelope features. This issue was evidenced in speaker identification results, where IF and GMR features added to the SEF vector produced a very modest improvement in the classification rate even though both the IF and GMR glottal feature sets had shown good discrimination ability when used alone. In light of this result, and given the large differences in predictability observed with respect to the particular spectral envelope feature-set used as the input to the GMR process (Appendix A.1, Tables 9 and 10), it is proposed that a straightforward way to incorporate glottal source information into a speech analysis application may be as simple as choosing an appropriate spectral envelope representation, whose level of glottal source content can be evaluated by the GMR approach presented in this thesis.

8.3 Contributions and Future Work

The contributions of this research consist of the following:

- A procedure for measuring and comparing the reliability of glottal inverse filtering algorithms and glottal waveform features that uses normalized measures of distance between adjacent estimates during segments of sustained phonation.
- The evaluation and comparison of reliability for four existing glottal inverse filtering methods on a large speech corpus. The set of studied IF methods

represents main ideas put forth in the inverse filtering literature and range from the classical approach of closed-phase inverse filtering to a recently proposed state-of-the-art procedure that is capable of modeling a time-varying vocal tract.

- The evaluation of measurement reliability and speaker identification ability for a set of widely studied glottal waveform parameters on a large speech corpus. The evaluated features included several alternatives for measuring the main characteristics of the glottal cycle.

- The development and evaluation of a statistical supervised learning procedure for estimating glottal waveform features directly from a spectral envelope feature vector, thus allowing for the characterization of the glottal airflow cycle without the need to perform inverse filtering.

- An objective assessment of the merit of using explicit glottal waveform information in speech analysis, over the alternative of implicitly using such information by performing speech analysis on conventional spectral envelope features.

Future work includes the evaluation of IF and GMR features on other speech analysis applications, such as the recognition of vocal affect. While the present results on speaker ID suggest only a small advantage due to the use of explicit glottal information over direct analysis with spectral envelope features (and the low reliability of IF features further suggests the generalizability of this result), the result should be confirmed in other speech applications by comparing the use of SEFs with and without IF and GMR features.

Furthermore, the results on measurement reliability of IF and GMR features suggest that the presented glottal feature estimation procedure may be applicable to clinical or forensic settings where an explicit assessment of glottal behavior is sought, as it allows an approximation to be obtained from the acoustic speech signal alone, but with a higher level of consistency than would result from an inverse filtering procedure

via current methods. The evaluation of the GMR procedure's applicability in this area would necessitate a comparison of the obtained GMR feature values to ground-truth information about glottal behavior. Such a comparison may be performed against glottal parameters manually extracted by clinicians from visual inspection of high-speed video of the vocal folds.

Finally, the availability of a large database containing ancillary signals (e.g. high-speed video) representing a more direct observation of glottal behavior from a large set of speakers may also enable training of the GMR model on much cleaner and reliable data than what can be obtained through IF methods. While the relationship between alternative signal modalities and glottal airflow is far from simple, and would become a component of this line of investigation, the trained GMMs may be able to capture relationships between glottal behavior and the spectral envelope of speech which cannot be consistently observed from glottal waveform estimates due to errors in the inverse filtering process.

APPENDIX A

SIMILARITY BETWEEN IF AND GMR FEATURES

A.1 Effect of Spectral Envelope Feature Set

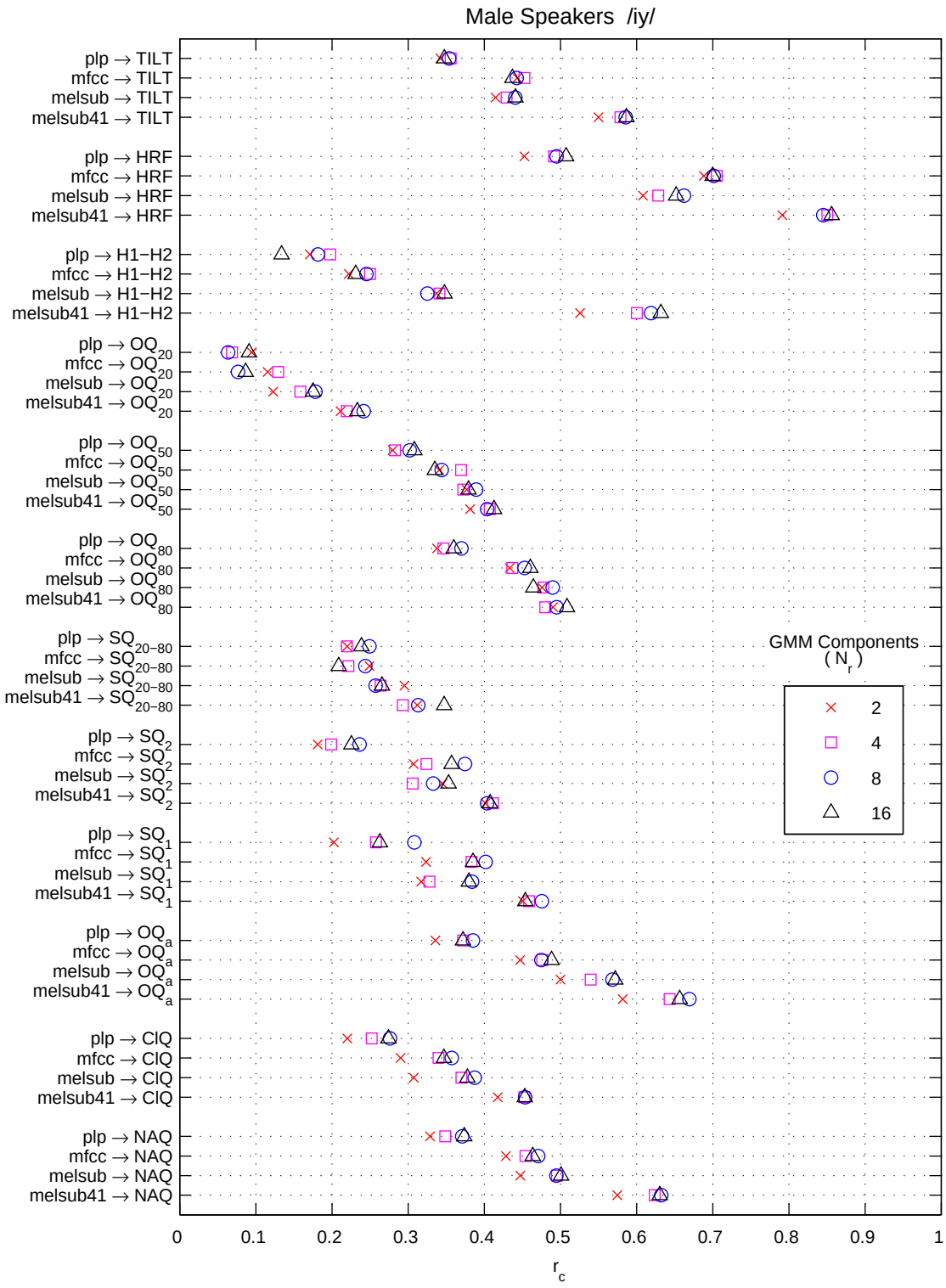


Figure 25: Correlation coefficient r_c between IF and GMR glottal features, Direct measurement features, phoneme /iy/, male speakers.

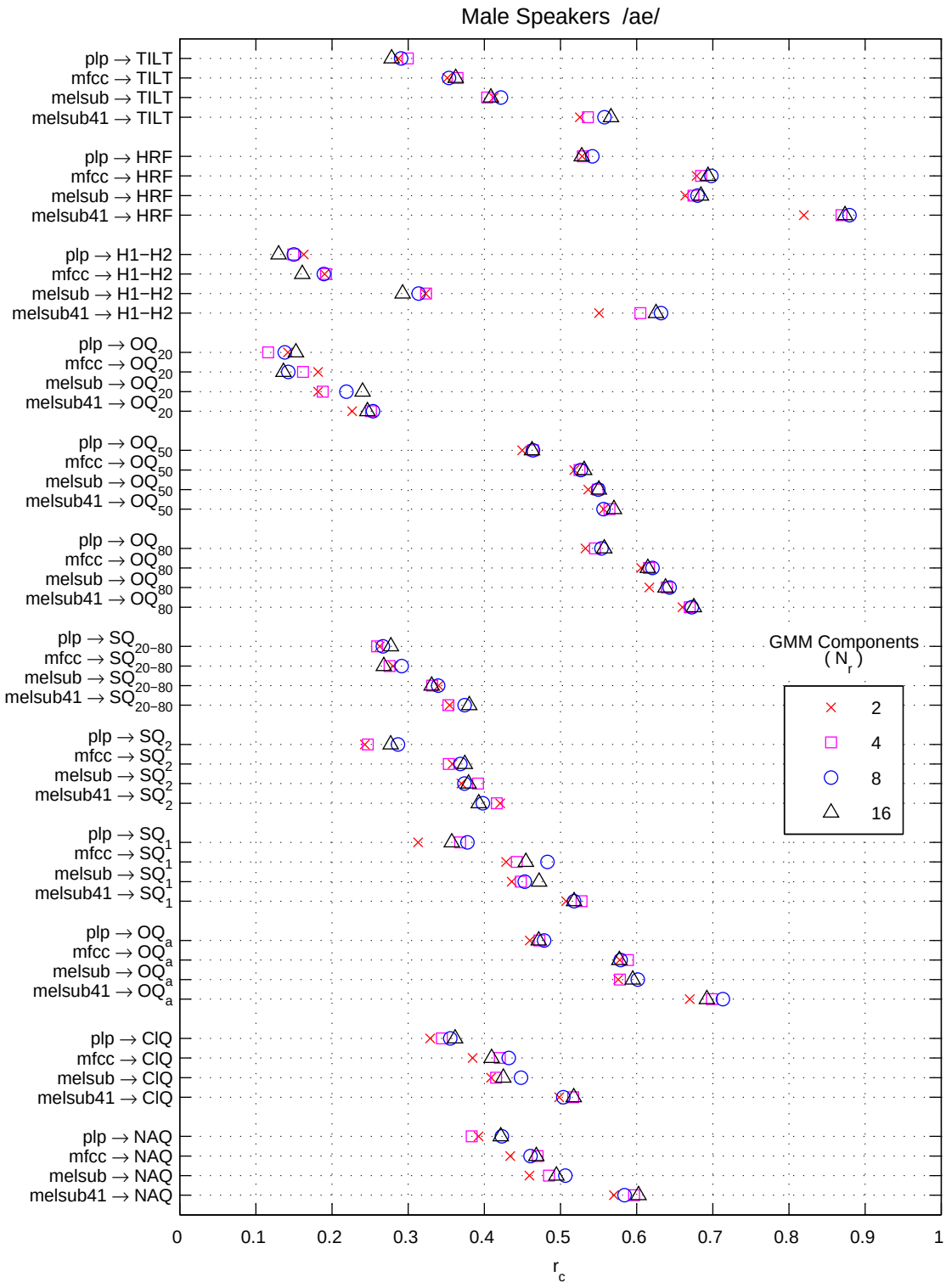


Figure 26: Correlation coefficient r_c between IF and GMR glottal features, Direct measurement features, phoneme /ae/, male speakers.

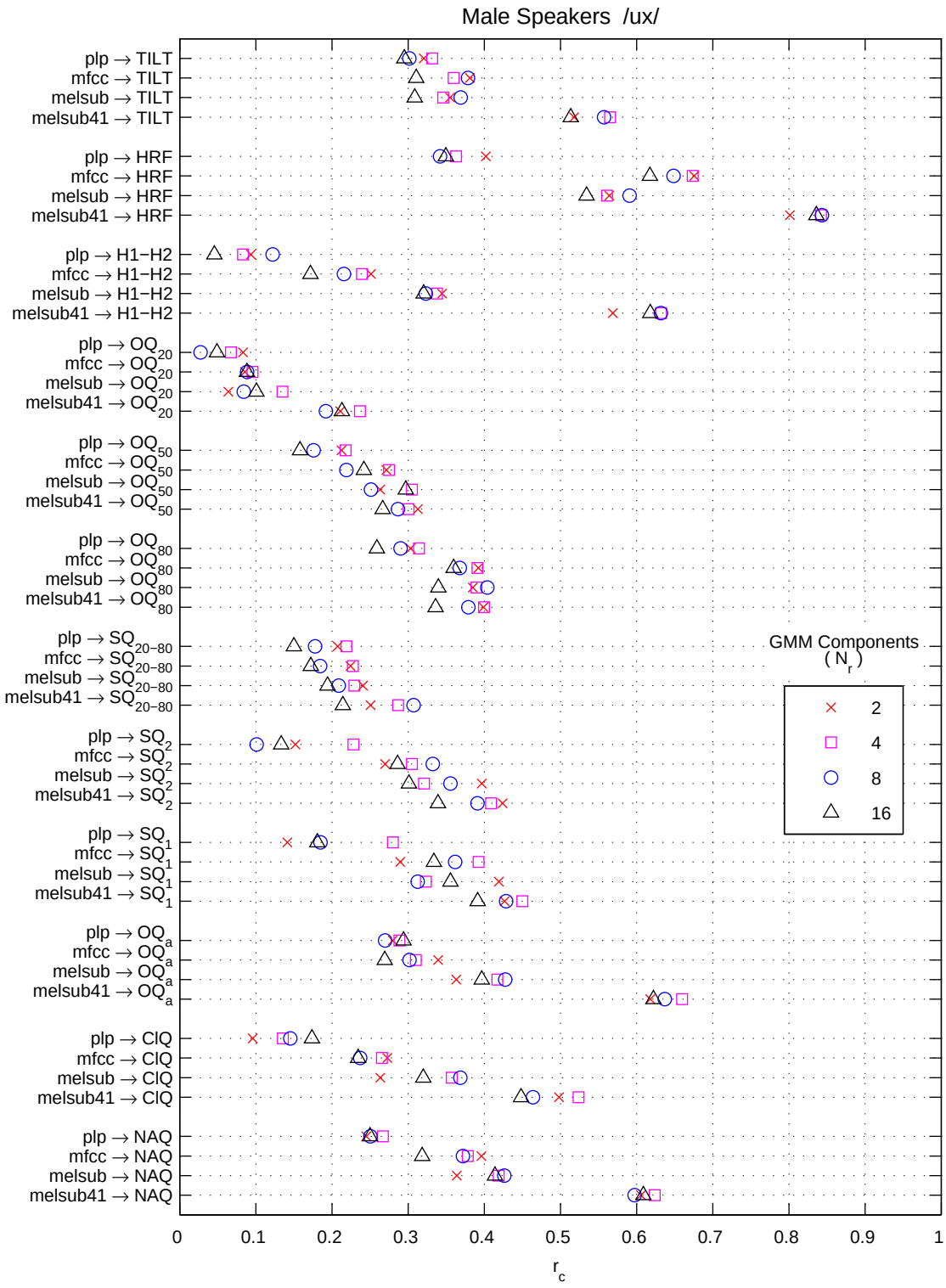


Figure 27: Correlation coefficient r_c between IF and GMR glottal features, Direct measurement features, phoneme /ux/, male speakers.

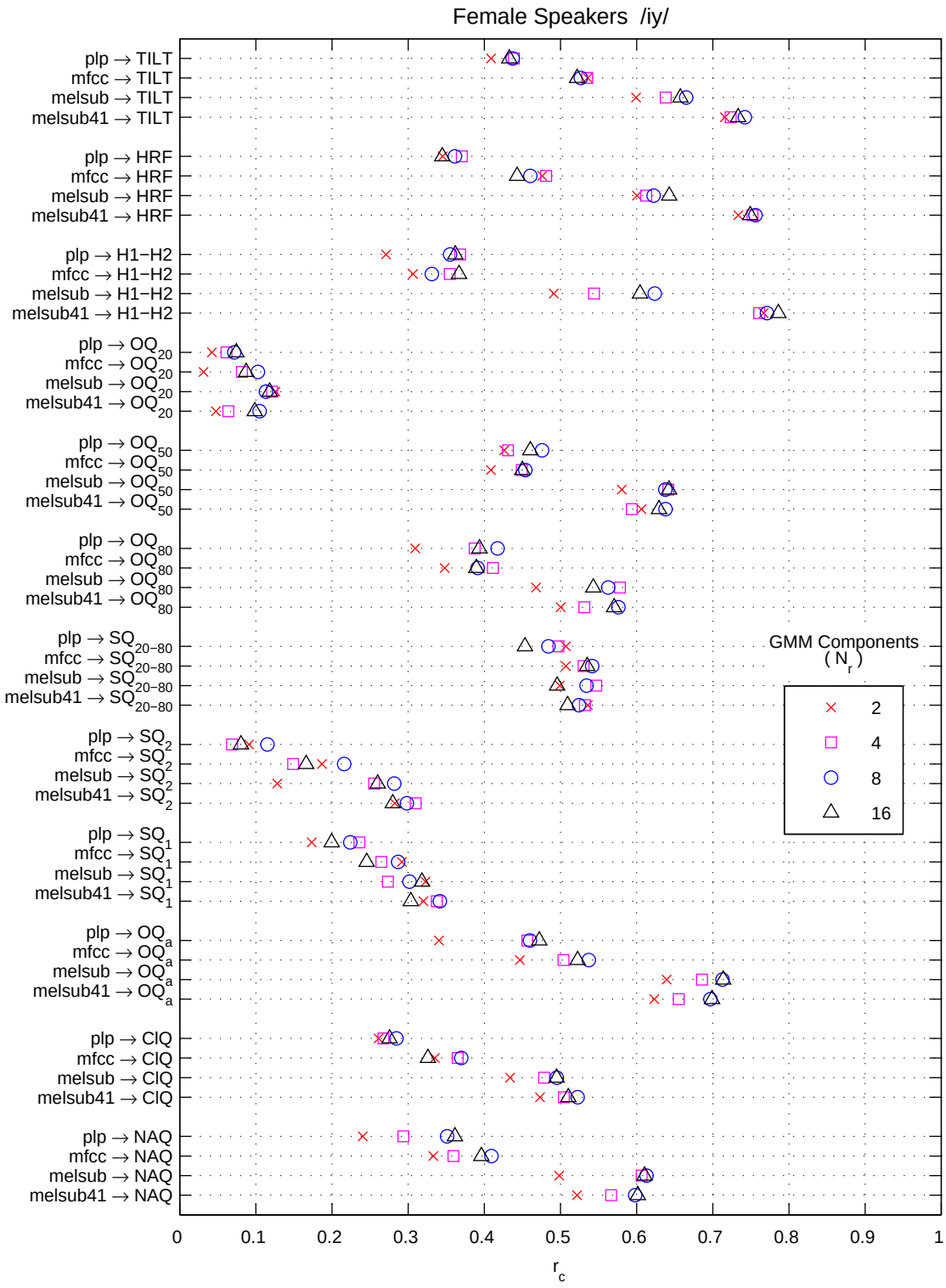


Figure 28: Correlation coefficient r_c between IF and GMR glottal features, Direct measurement features, phoneme /iy/, female speakers.

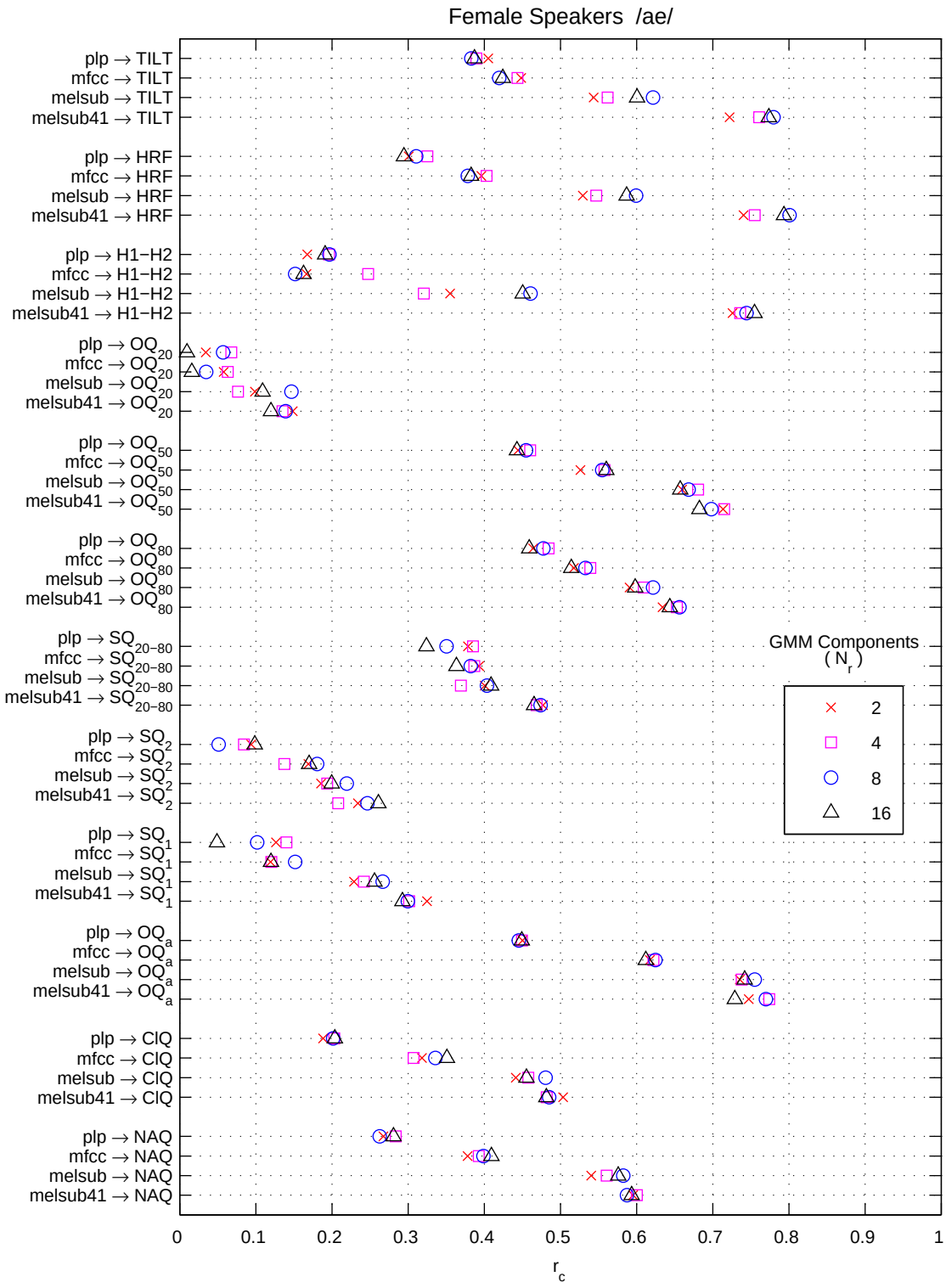


Figure 29: Correlation coefficient r_c between IF and GMR glottal features, Direct measurement features, phoneme /ae/, female speakers.

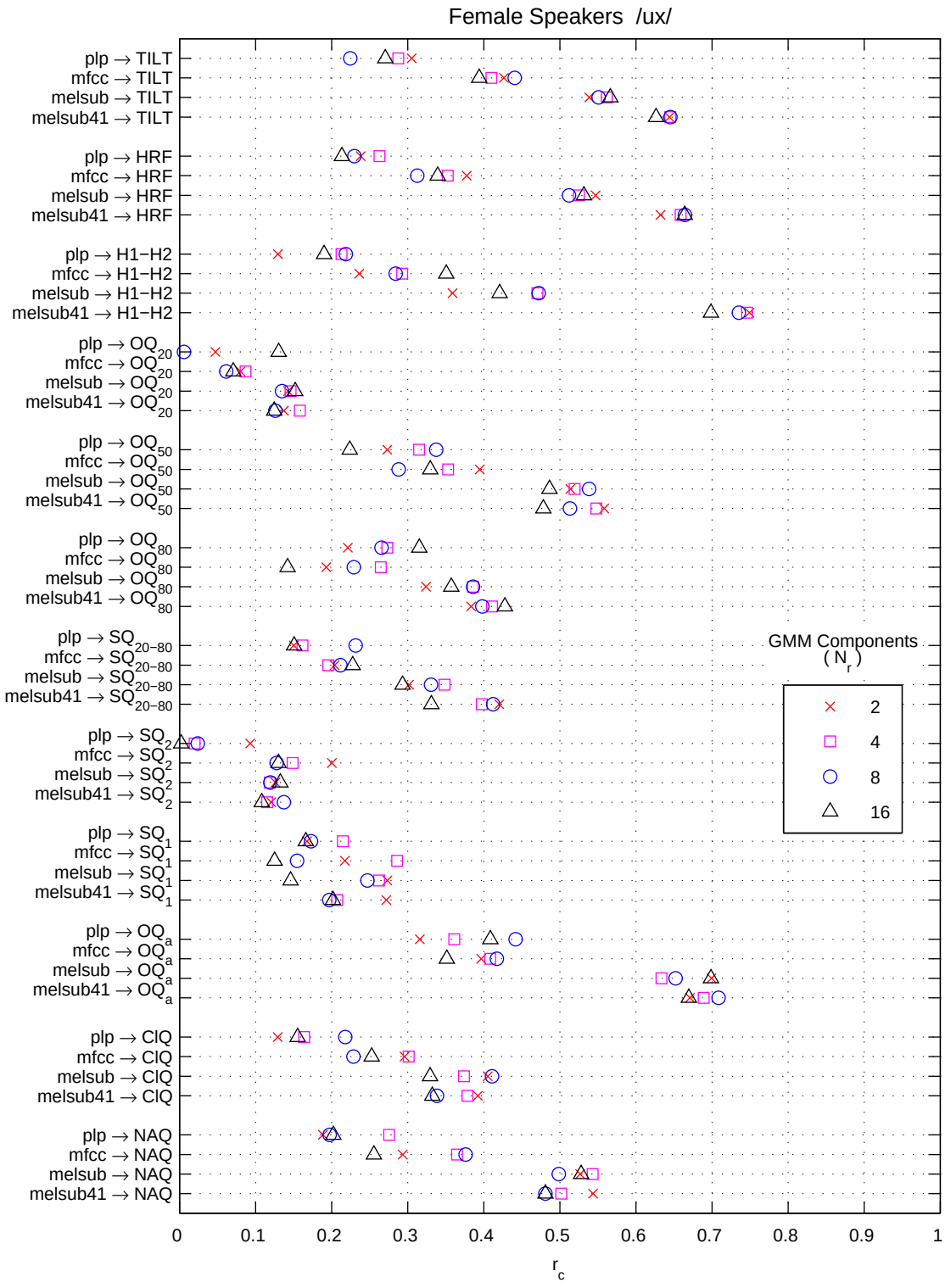


Figure 30: Correlation coefficient r_c between IF and GMR glottal features, Direct measurement features, phoneme /ux/, female speakers.

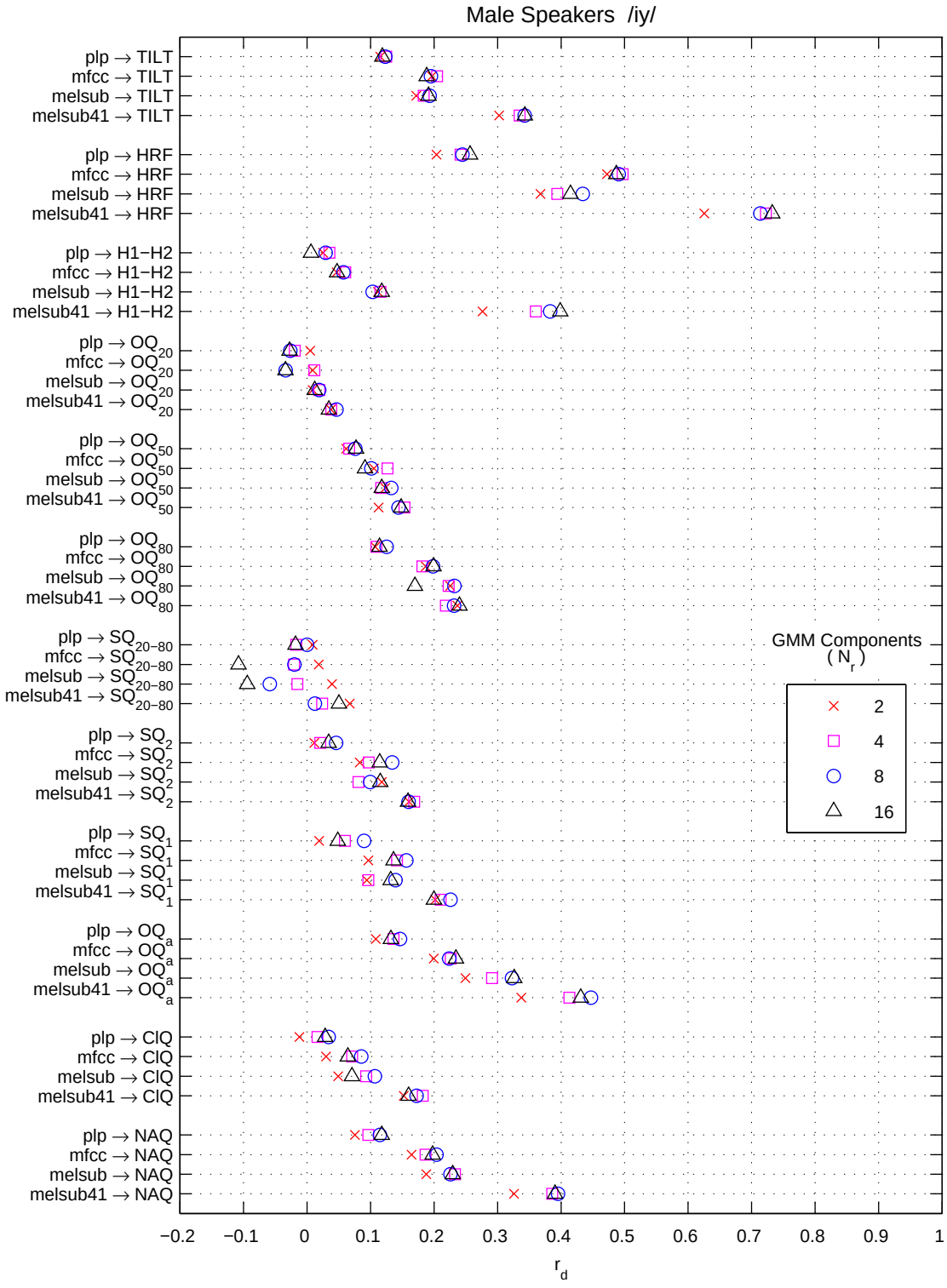


Figure 31: Coefficient of determination r_d between IF and GMR glottal features, Direct measurement features, phoneme /iy/, male speakers.

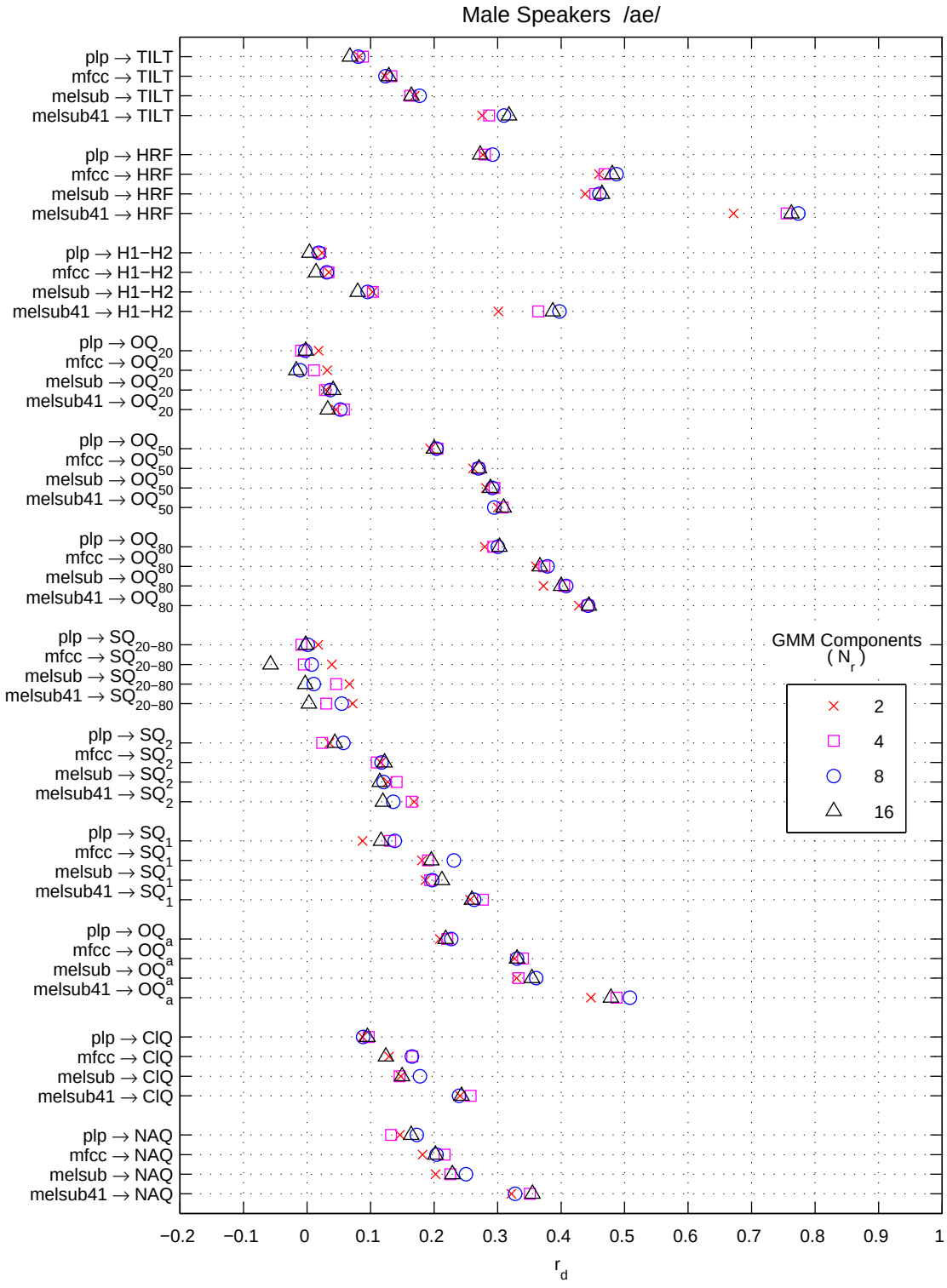


Figure 32: Coefficient of determination r_d between IF and GMR glottal features, Direct measurement features, phoneme /ae/, male speakers.

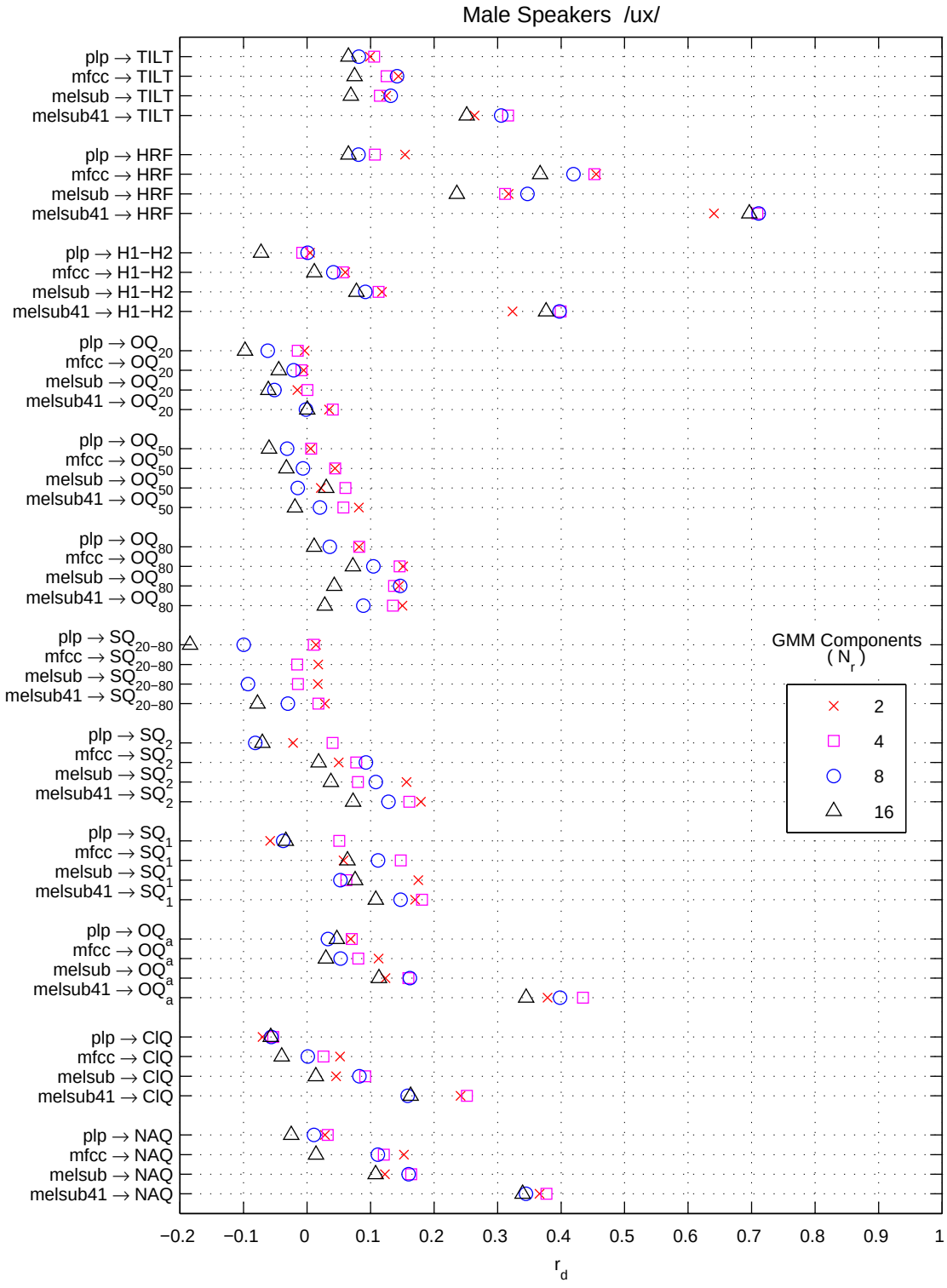


Figure 33: Coefficient of determination r_d between IF and GMR glottal features, Direct measurement features, phoneme /ux/, male speakers.

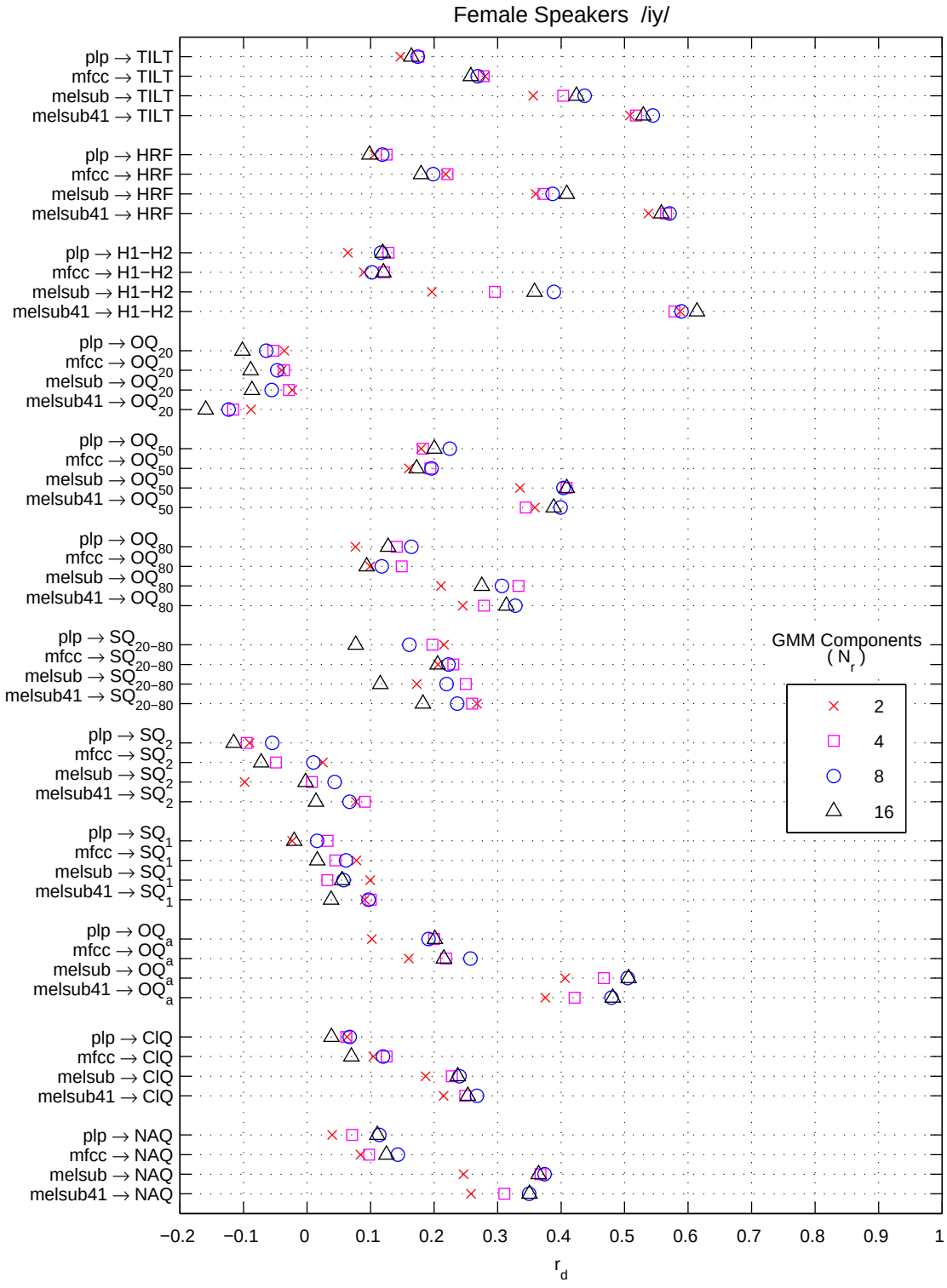


Figure 34: Coefficient of determination r_d between IF and GMR glottal features, Direct measurement features, phoneme /iy/, female speakers.

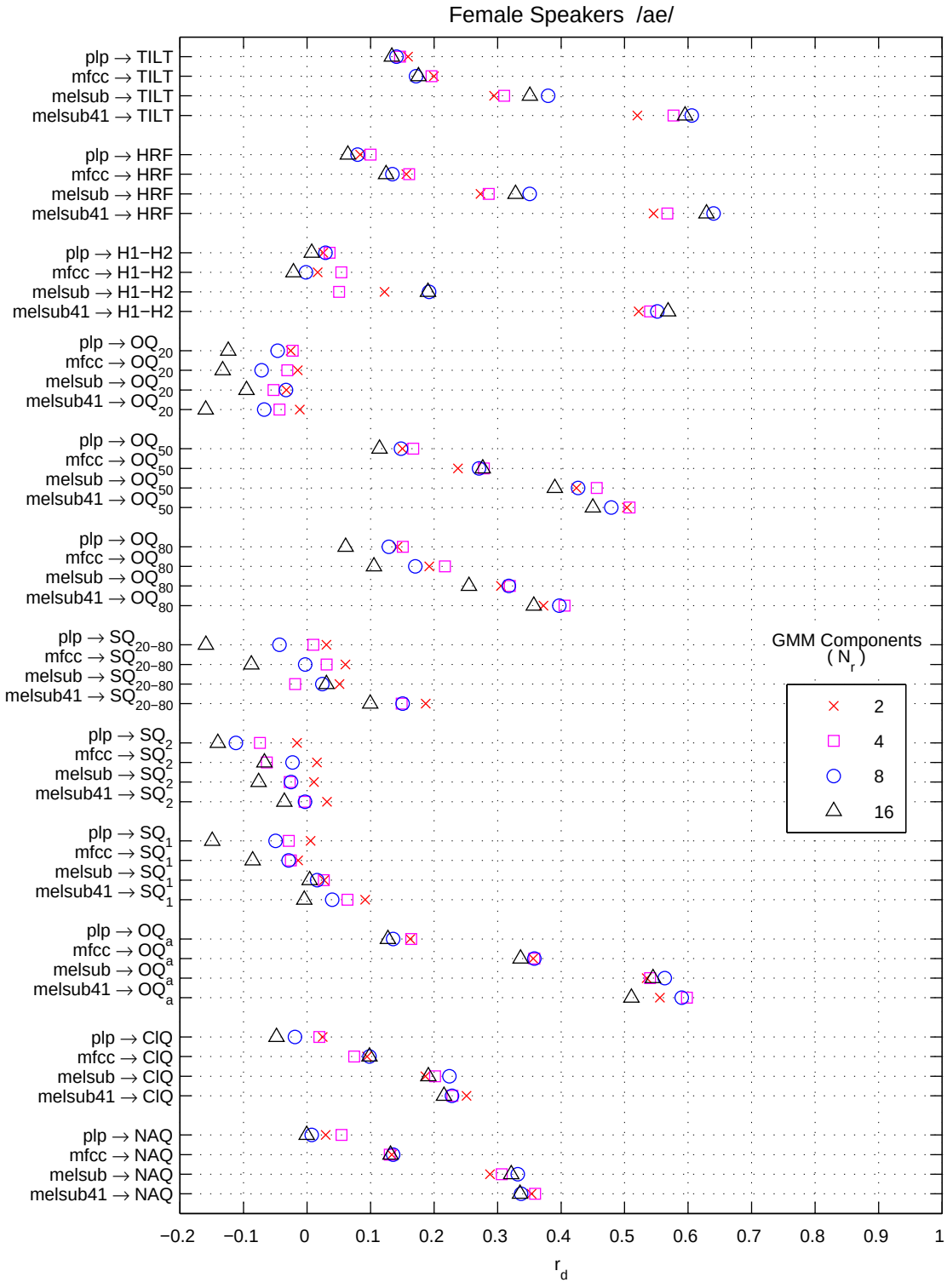


Figure 35: Coefficient of determination r_d between IF and GMR glottal features, Direct measurement features, phoneme /ae/, female speakers.

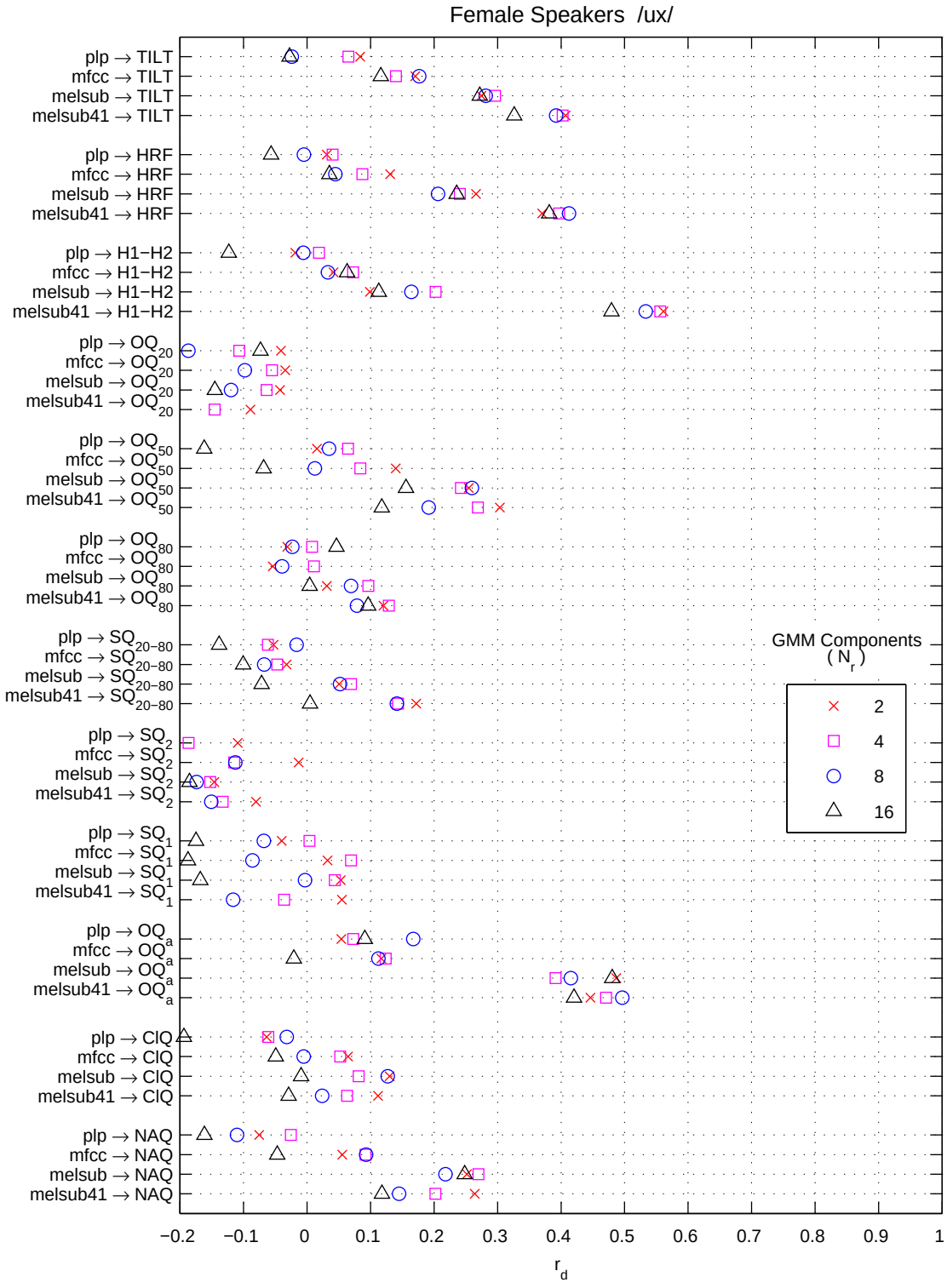


Figure 36: Coefficient of determination r_d between IF and GMR glottal features, Direct measurement features, phoneme /ux/, female speakers.

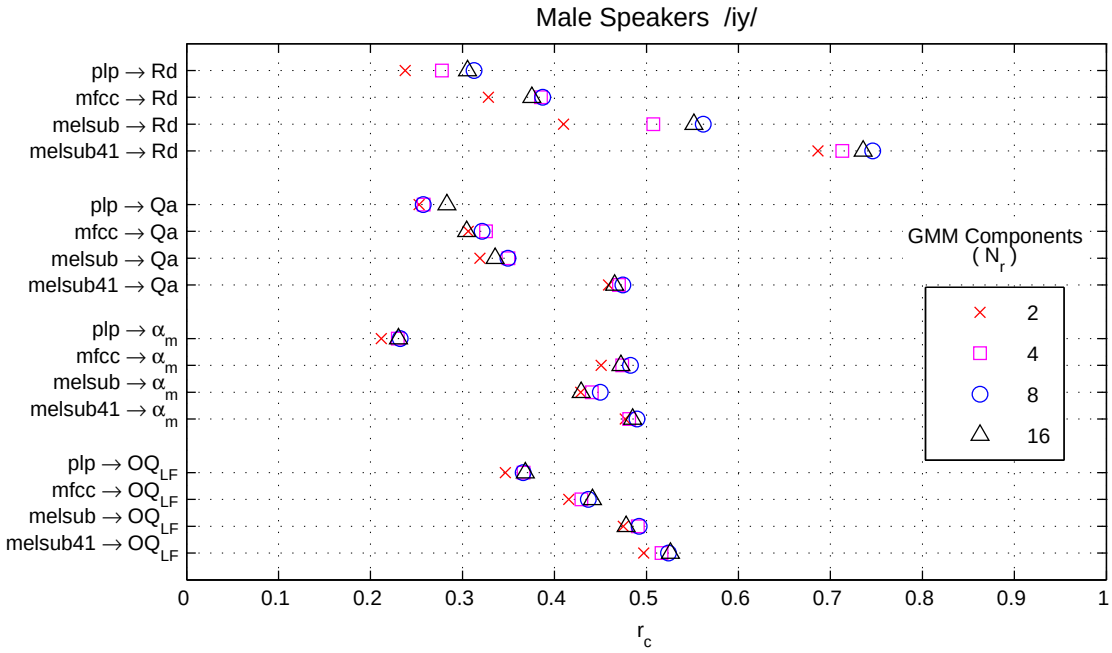


Figure 37: Correlation coefficient r_c between IF and GMR glottal features, LF-model features, phoneme /iy/, male speakers.

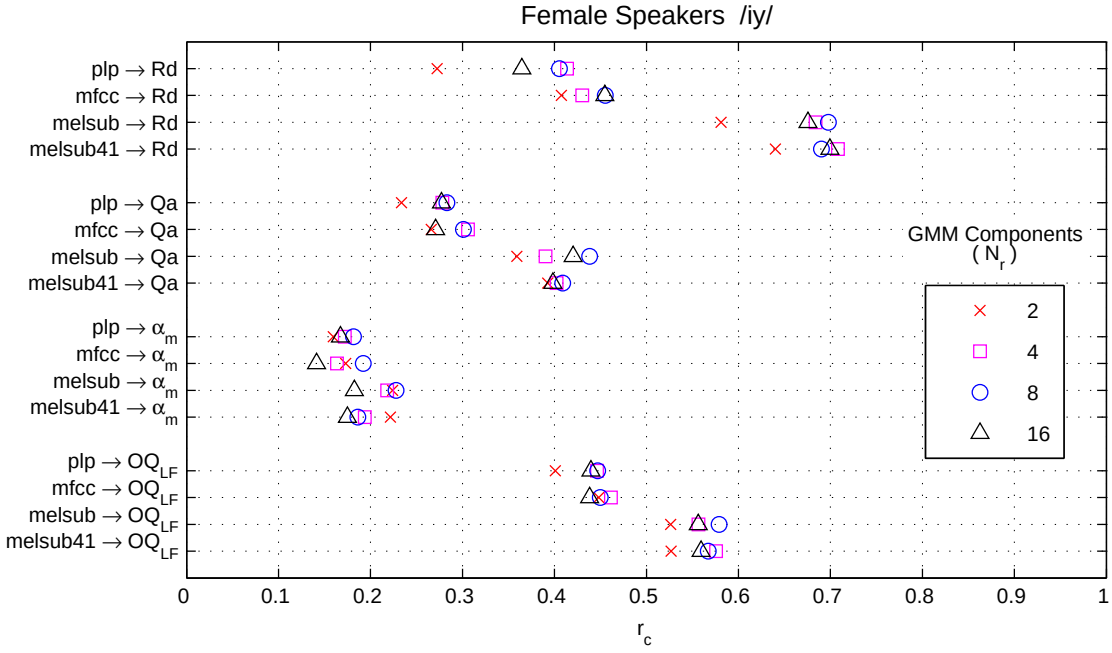


Figure 38: Correlation coefficient r_c between IF and GMR glottal features, LF-model features, phoneme /iy/, female speakers.

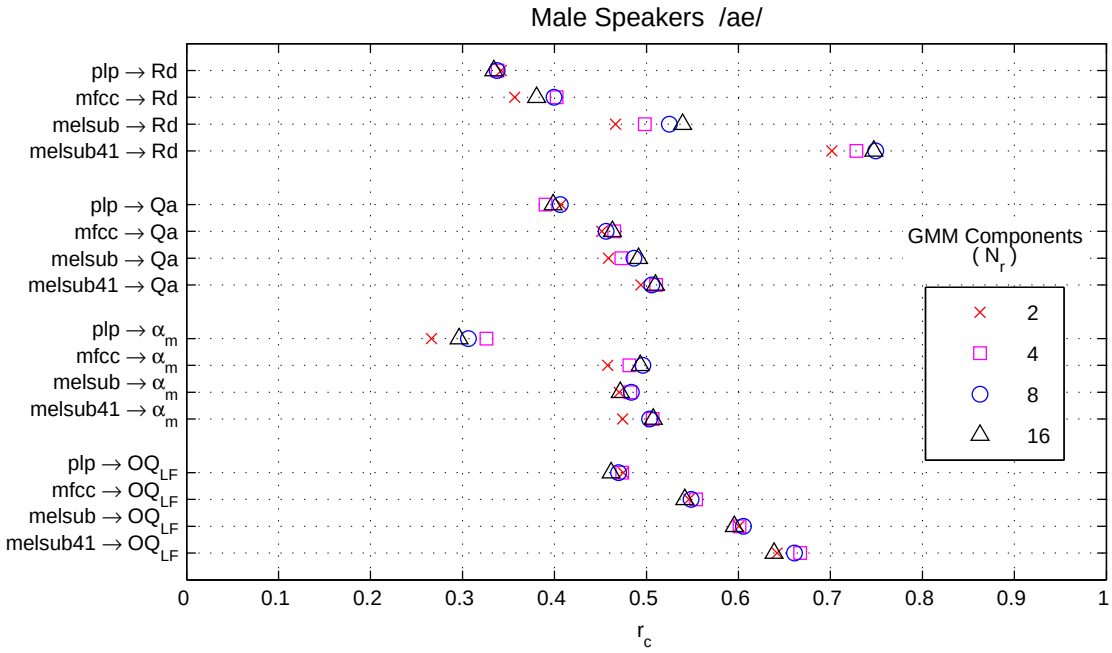


Figure 39: Correlation coefficient r_c between IF and GMR glottal features, LF-model features, phoneme /ae/, male speakers.

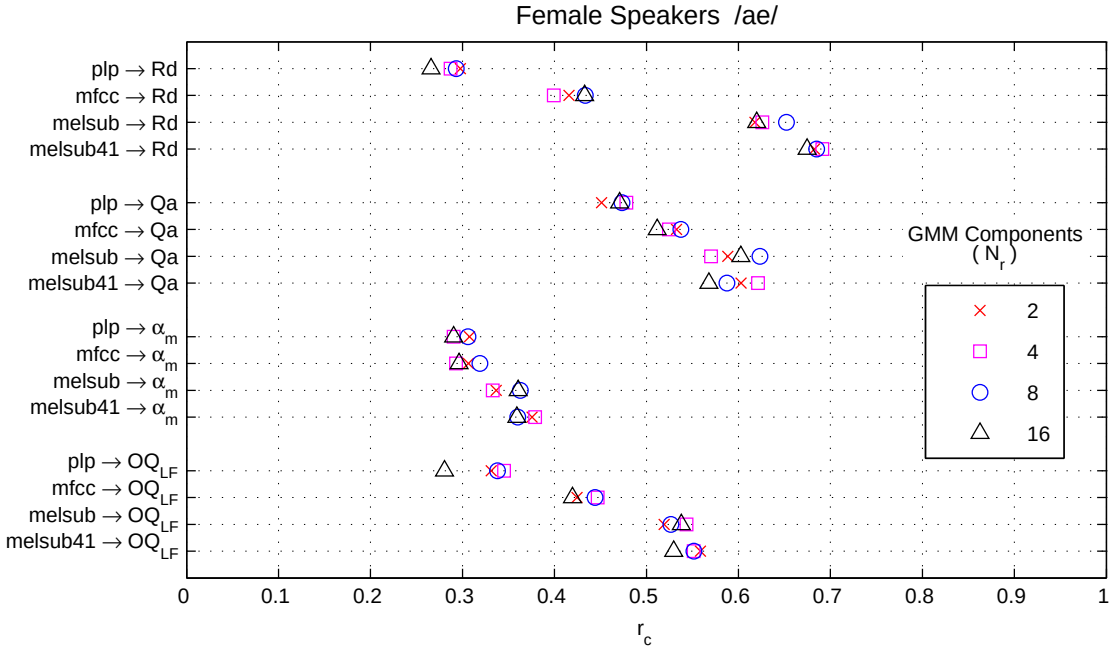


Figure 40: Correlation coefficient r_c between IF and GMR glottal features, LF-model features, phoneme /ae/, female speakers.

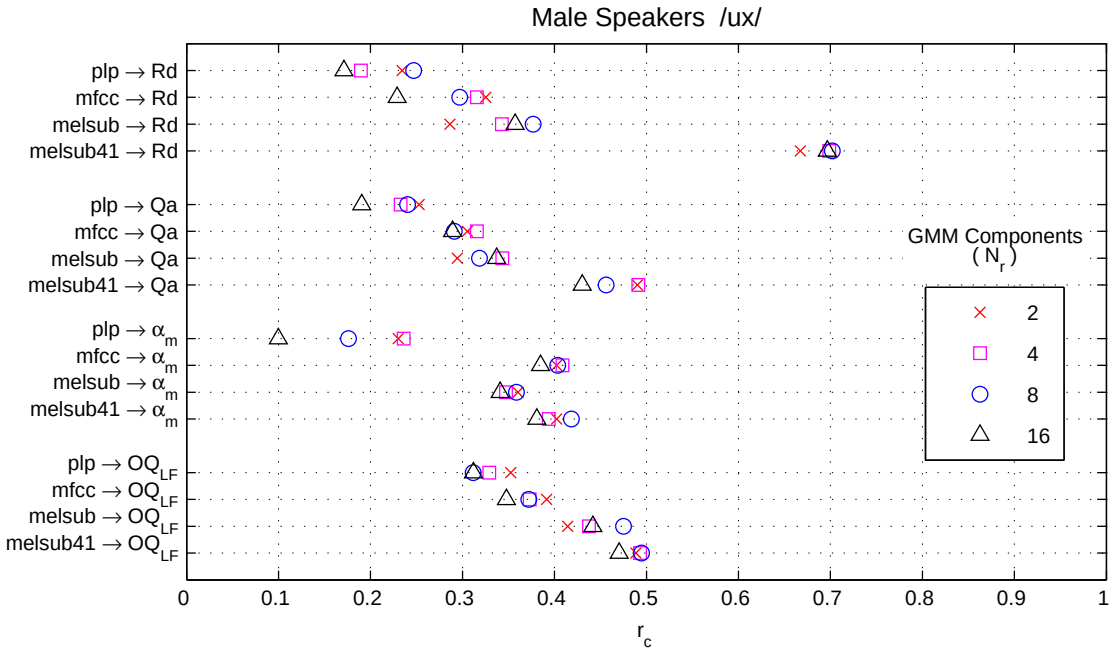


Figure 41: Correlation coefficient r_c between IF and GMR glottal features, LF-model features, phoneme /ux/, male speakers.

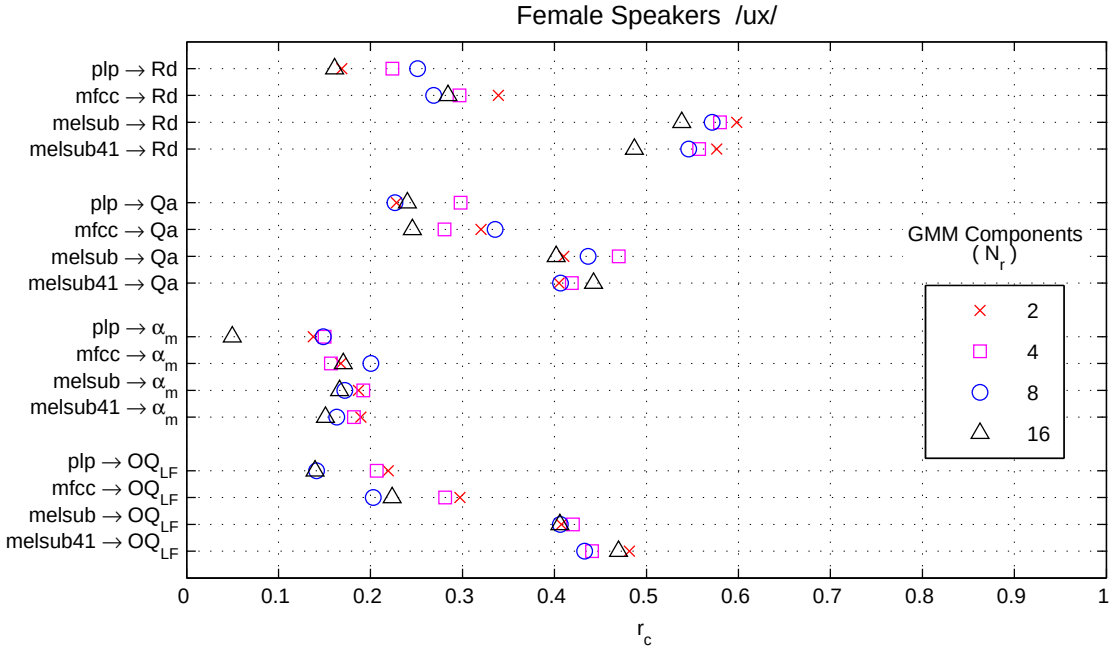


Figure 42: Correlation coefficient r_c between IF and GMR glottal features, LF-model features, phoneme /ux/, female speakers.

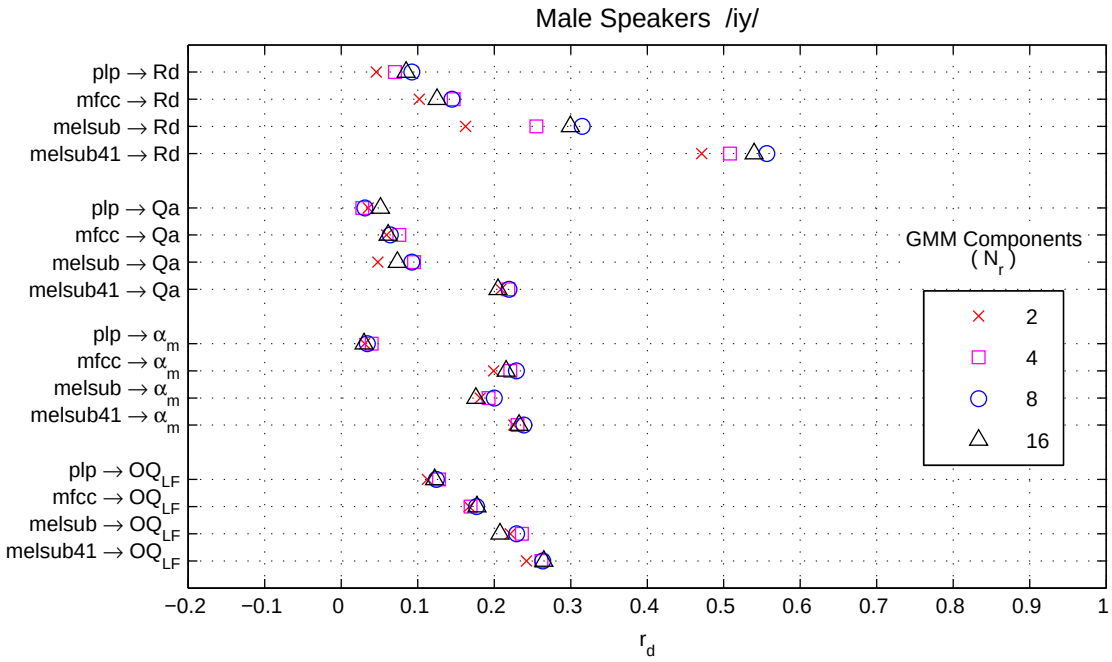


Figure 43: Coefficient of determination r_d between IF and GMR glottal features, LF-model features, phoneme /iy/, male speakers.

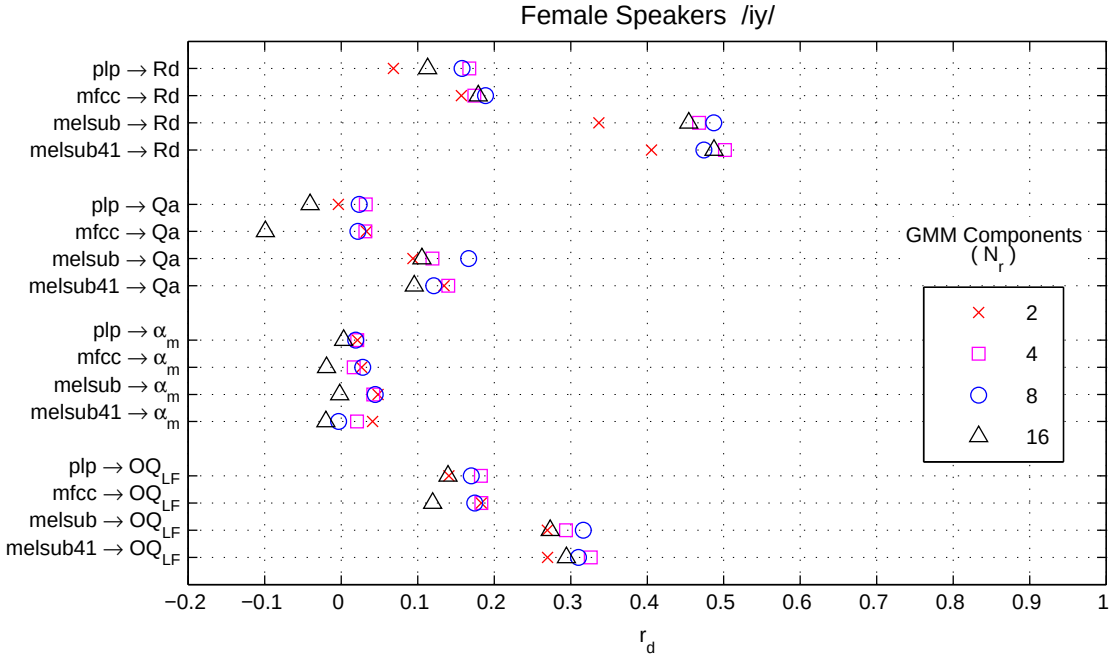


Figure 44: Coefficient of determination r_d between IF and GMR glottal features, LF-model features, phoneme /iy/, female speakers.

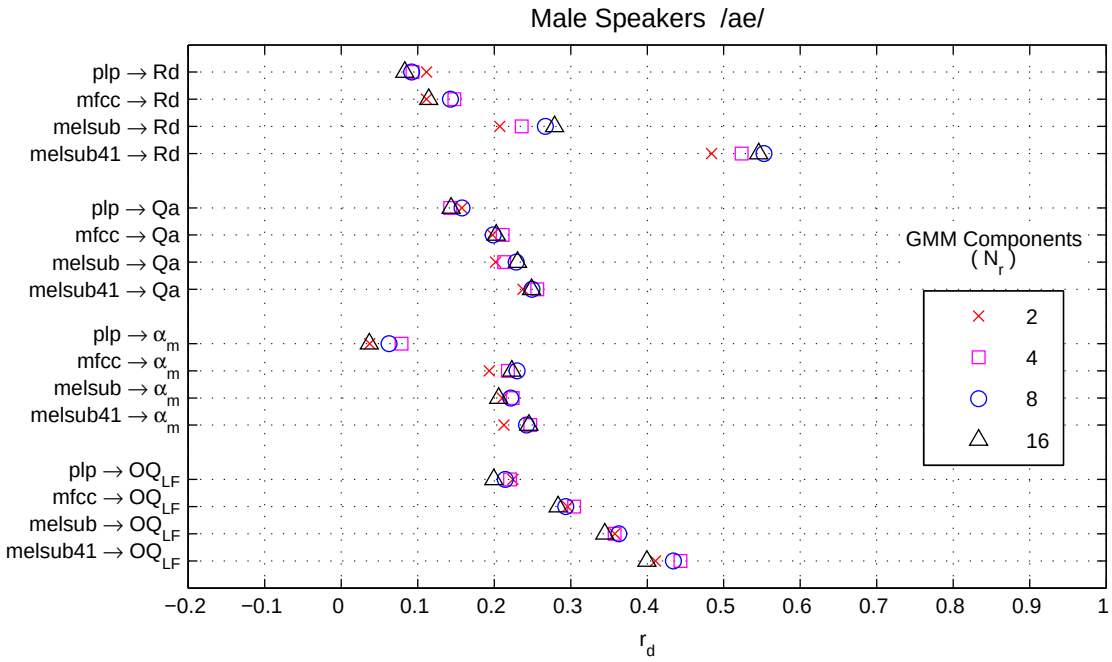


Figure 45: Coefficient of determination r_d between IF and GMR glottal features, LF-model features, phoneme /ae/, male speakers.

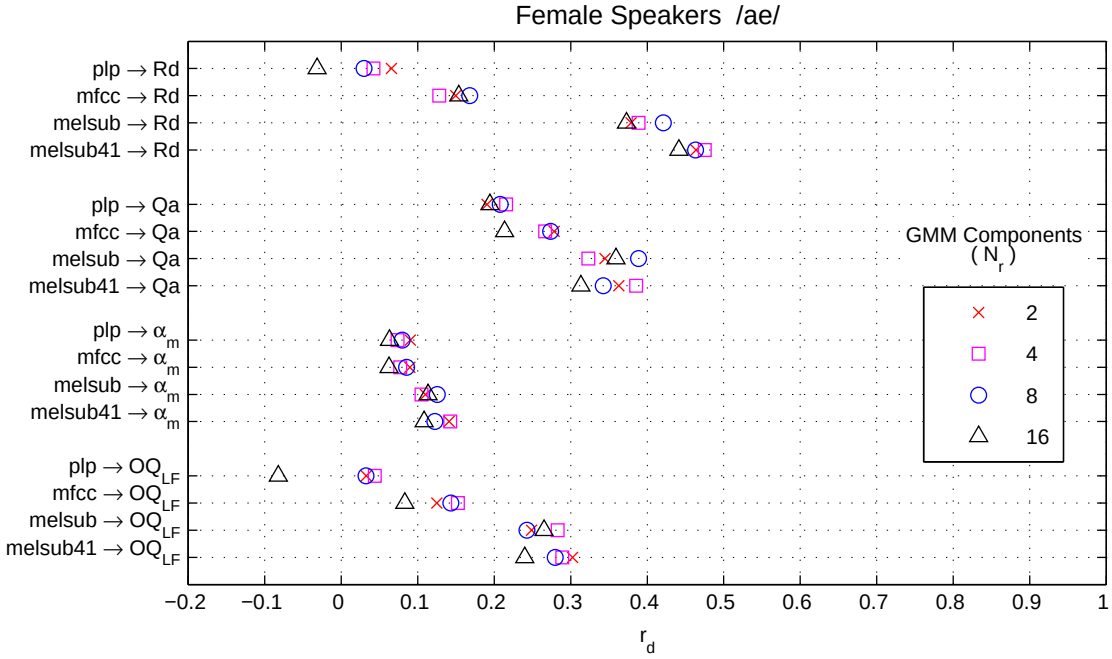


Figure 46: Coefficient of determination r_d between IF and GMR glottal features, LF-model features, phoneme /ae/, female speakers.

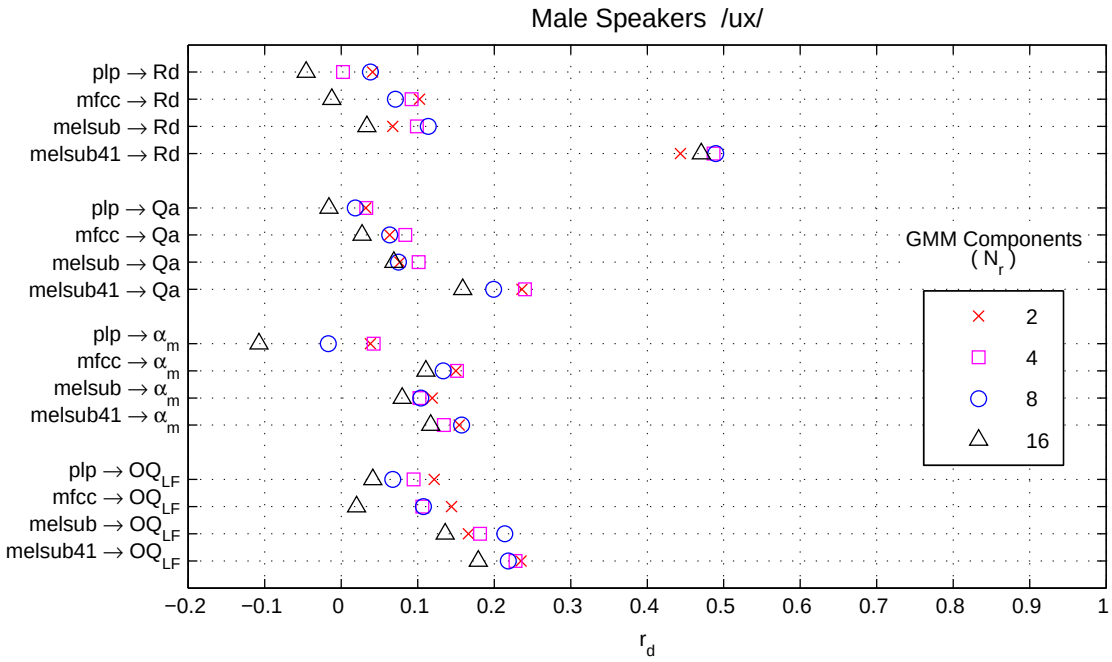


Figure 47: Coefficient of determination r_d between IF and GMR glottal features, LF-model features, phoneme /ux/, male speakers.

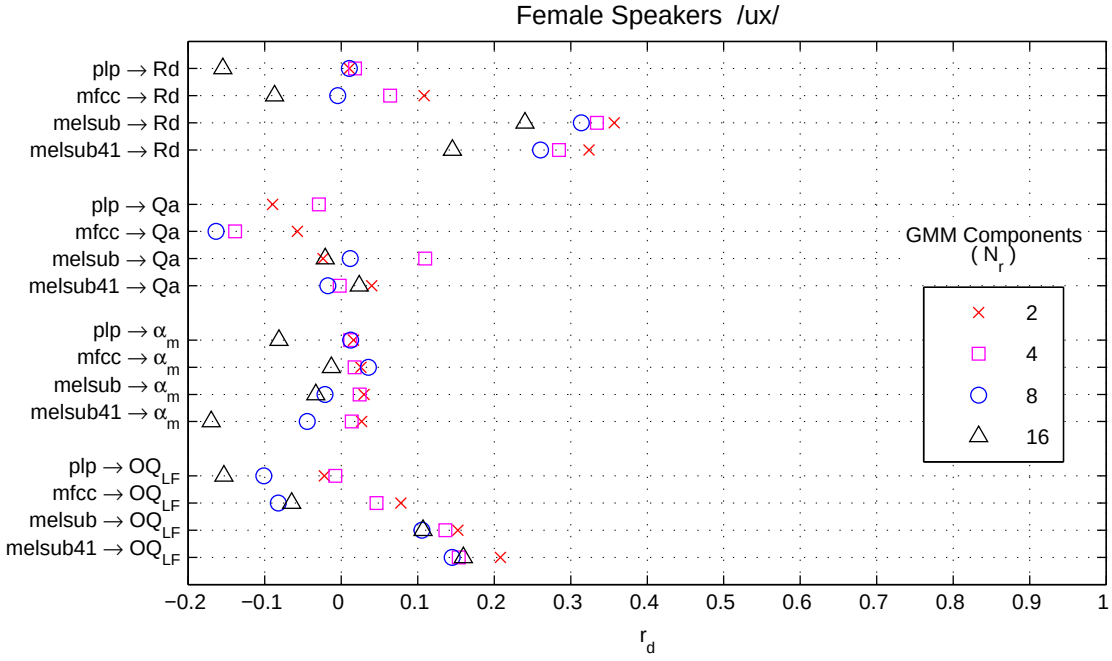
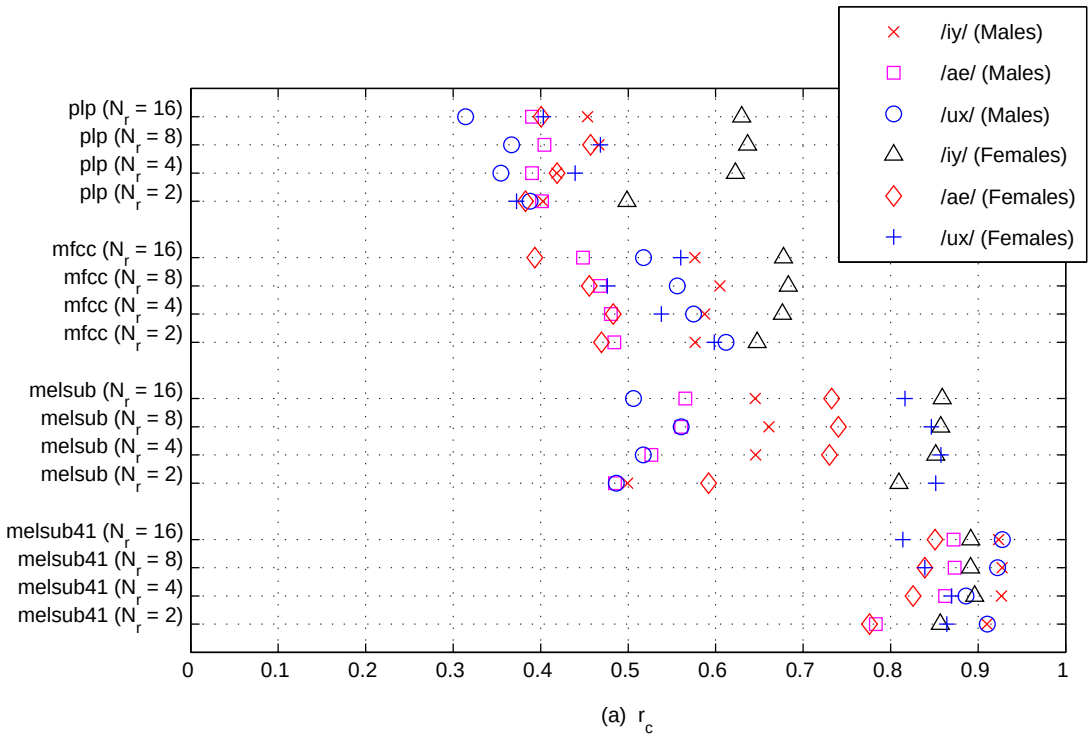


Figure 48: Coefficient of determination r_d between IF and GMR glottal features, LF-model features, phoneme /ux/, female speakers.



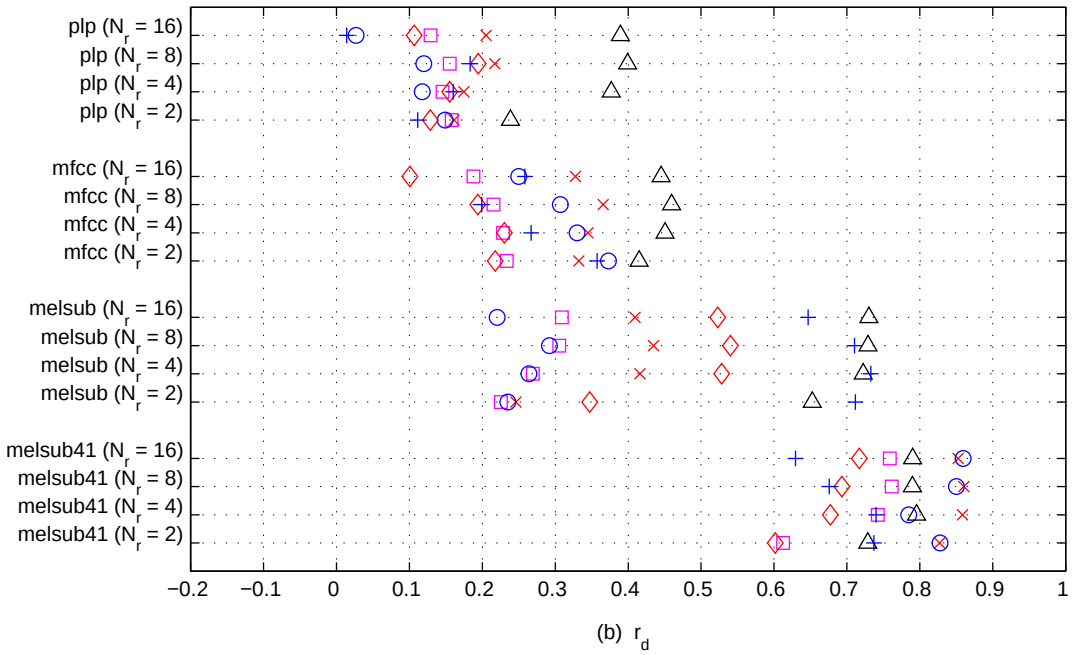


Figure 49: Correlation coefficient r_c (a) and coefficient of determination r_d (b) between RAPT and GMR pitch estimates (f_0) for each choice of SEF and number of GMM components (N_r).

A.2 Training on Pitch and Delta Features

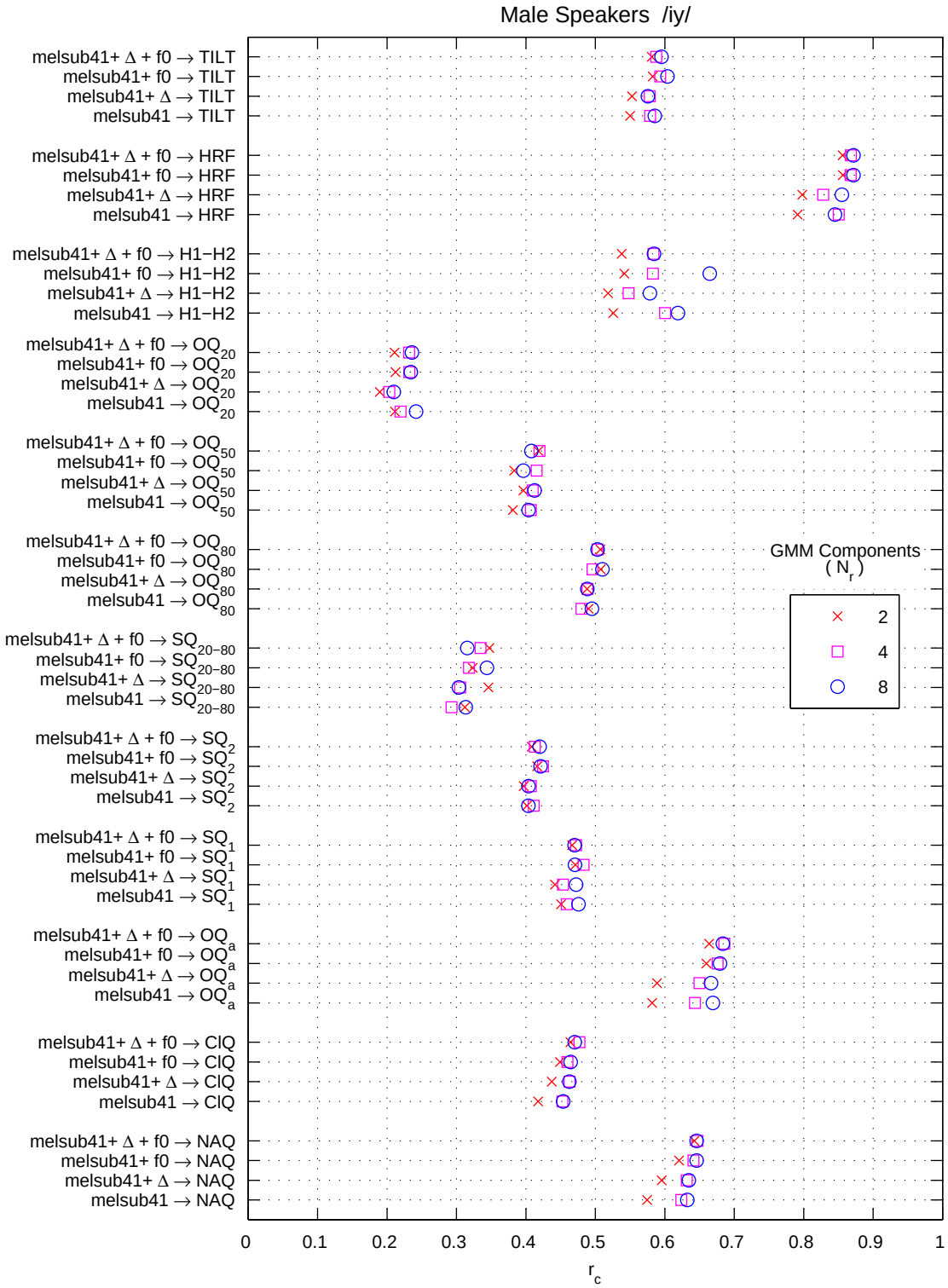


Figure 50: Correlation coefficient r_c between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /iy/, male speakers.

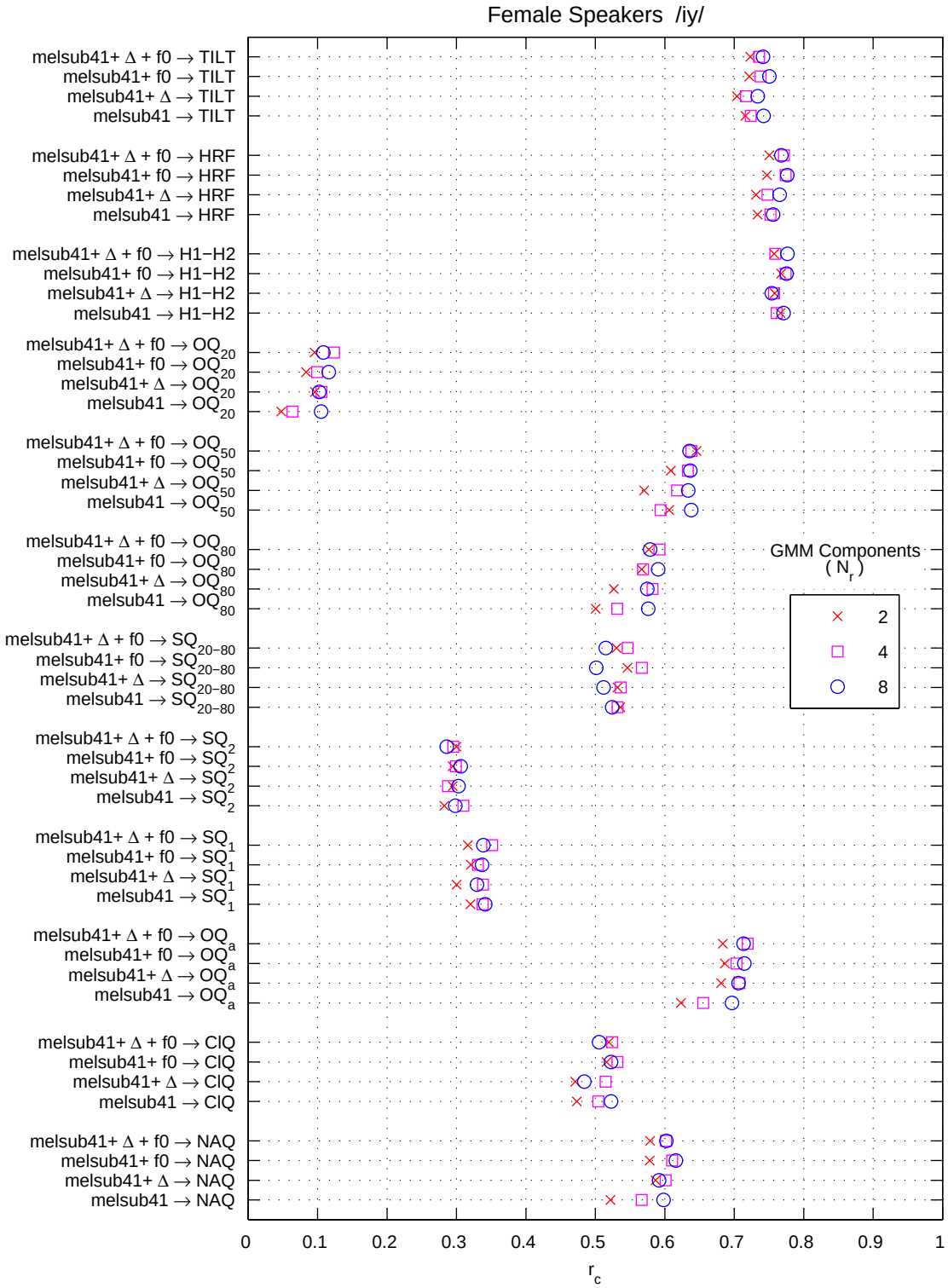


Figure 51: Correlation coefficient r_c between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /iy/, female speakers.

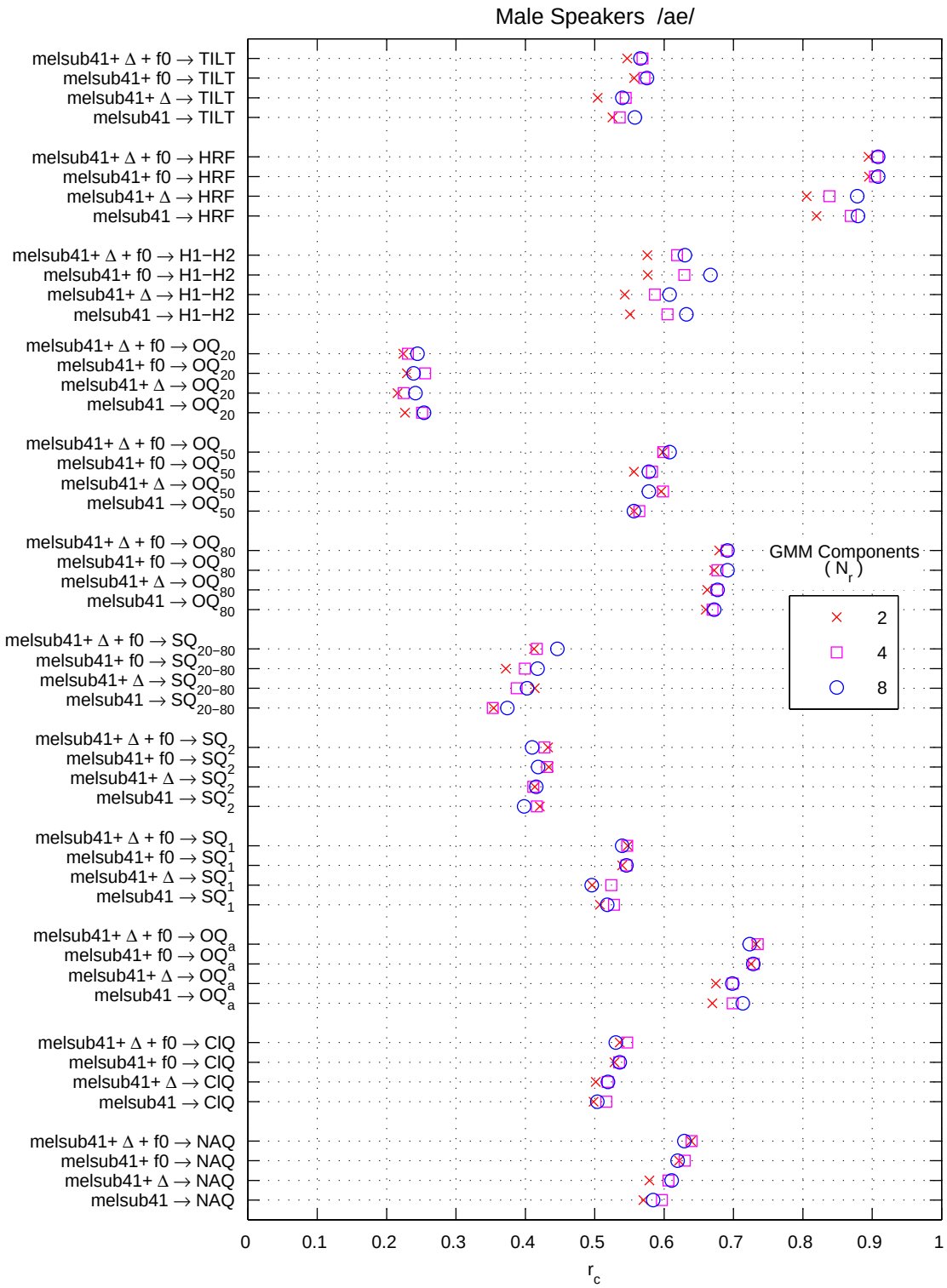


Figure 52: Correlation coefficient r_c between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ae/, male speakers.

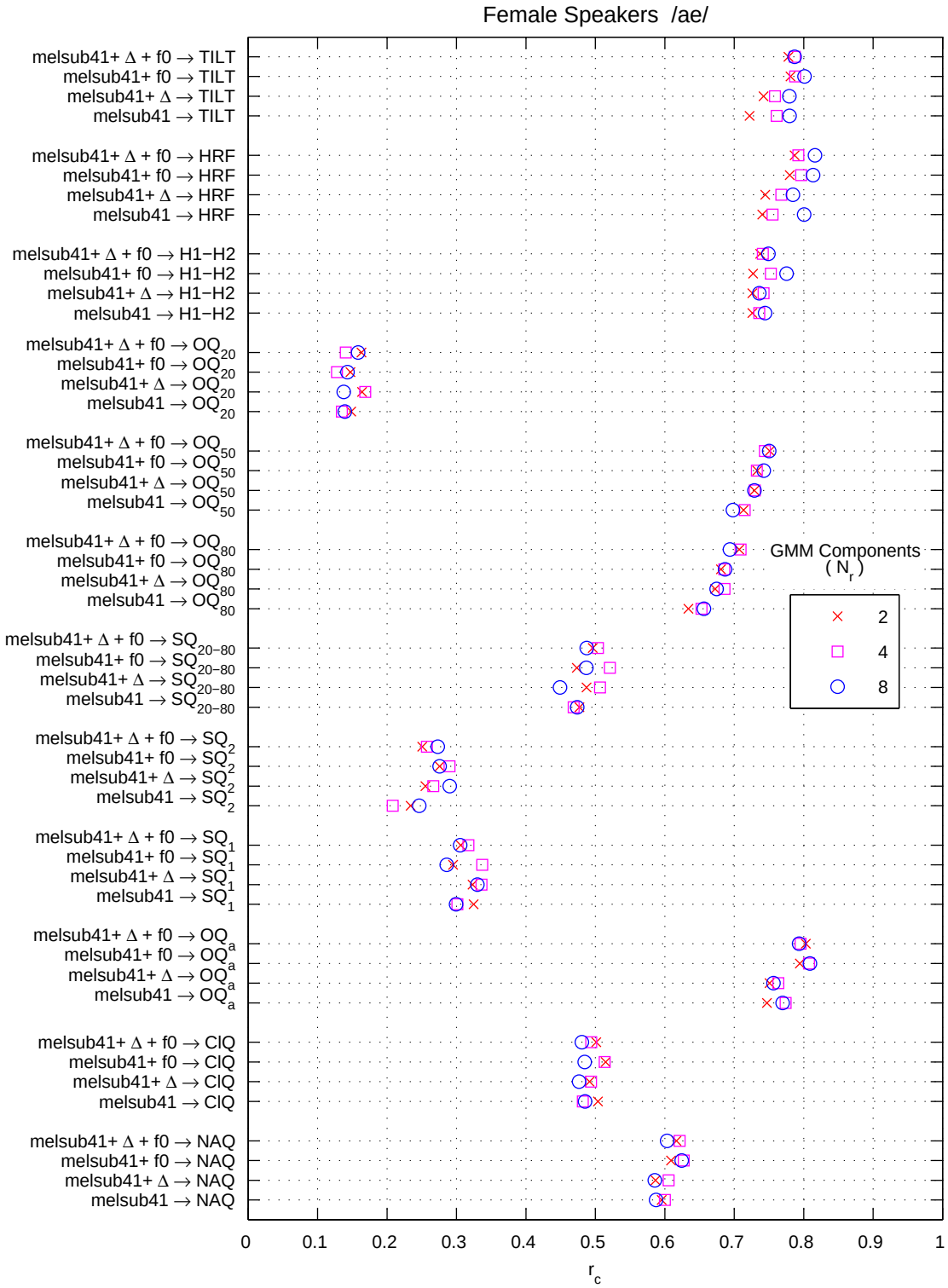


Figure 53: Correlation coefficient r_c between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ae/, female speakers.

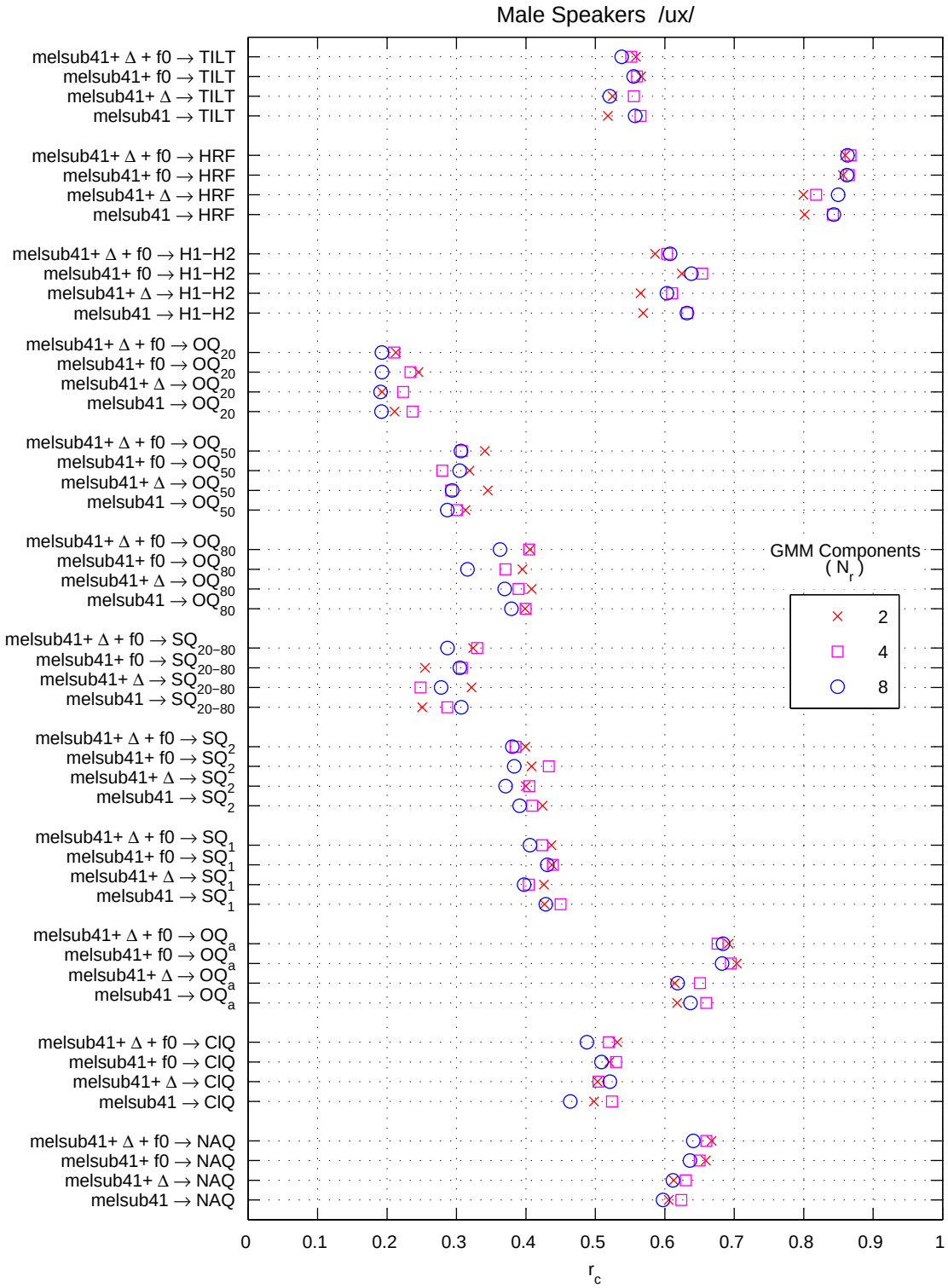


Figure 54: Correlation coefficient r_c between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ux/, male speakers.

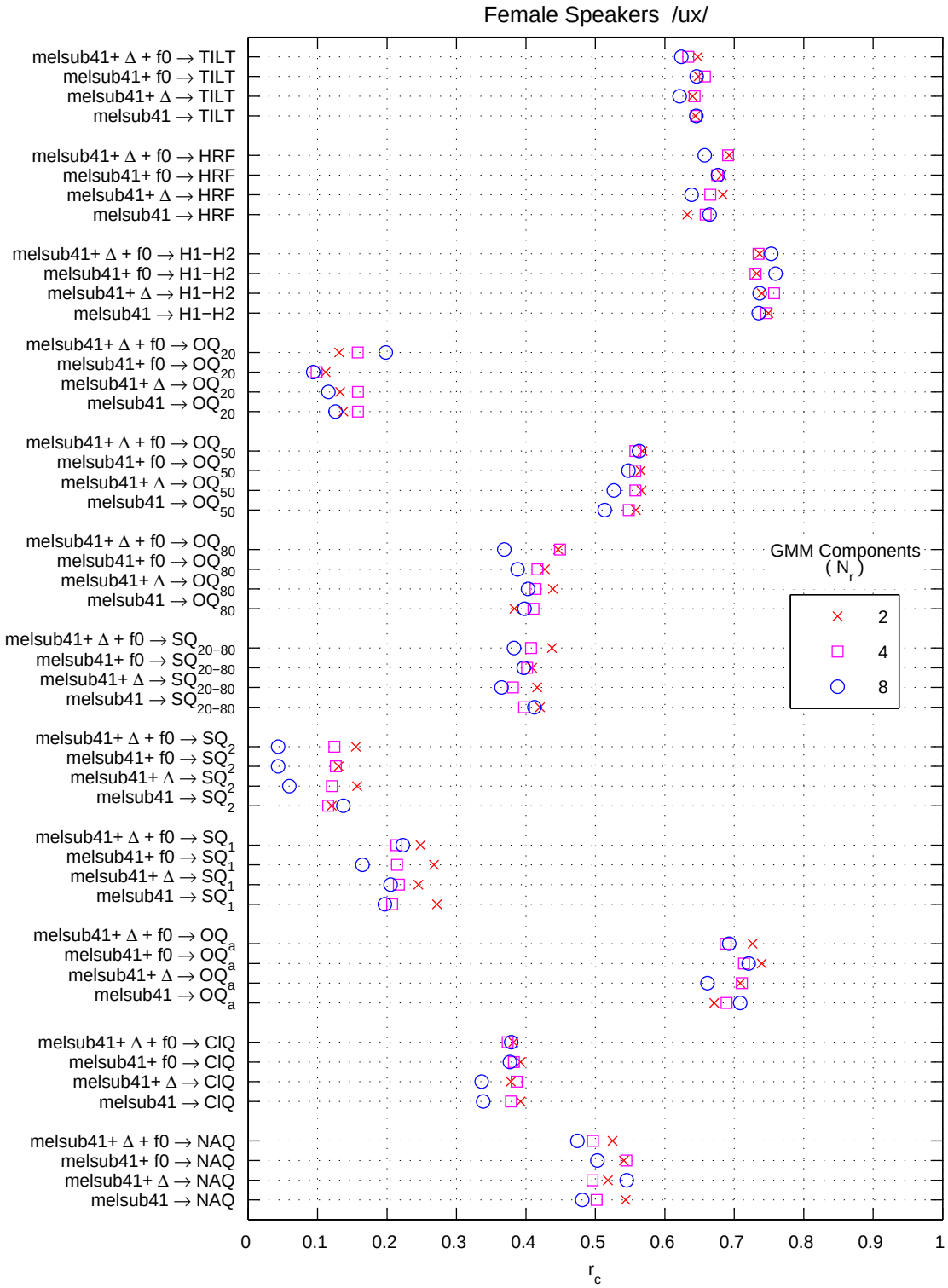


Figure 55: Correlation coefficient r_c between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ux/, female speakers.

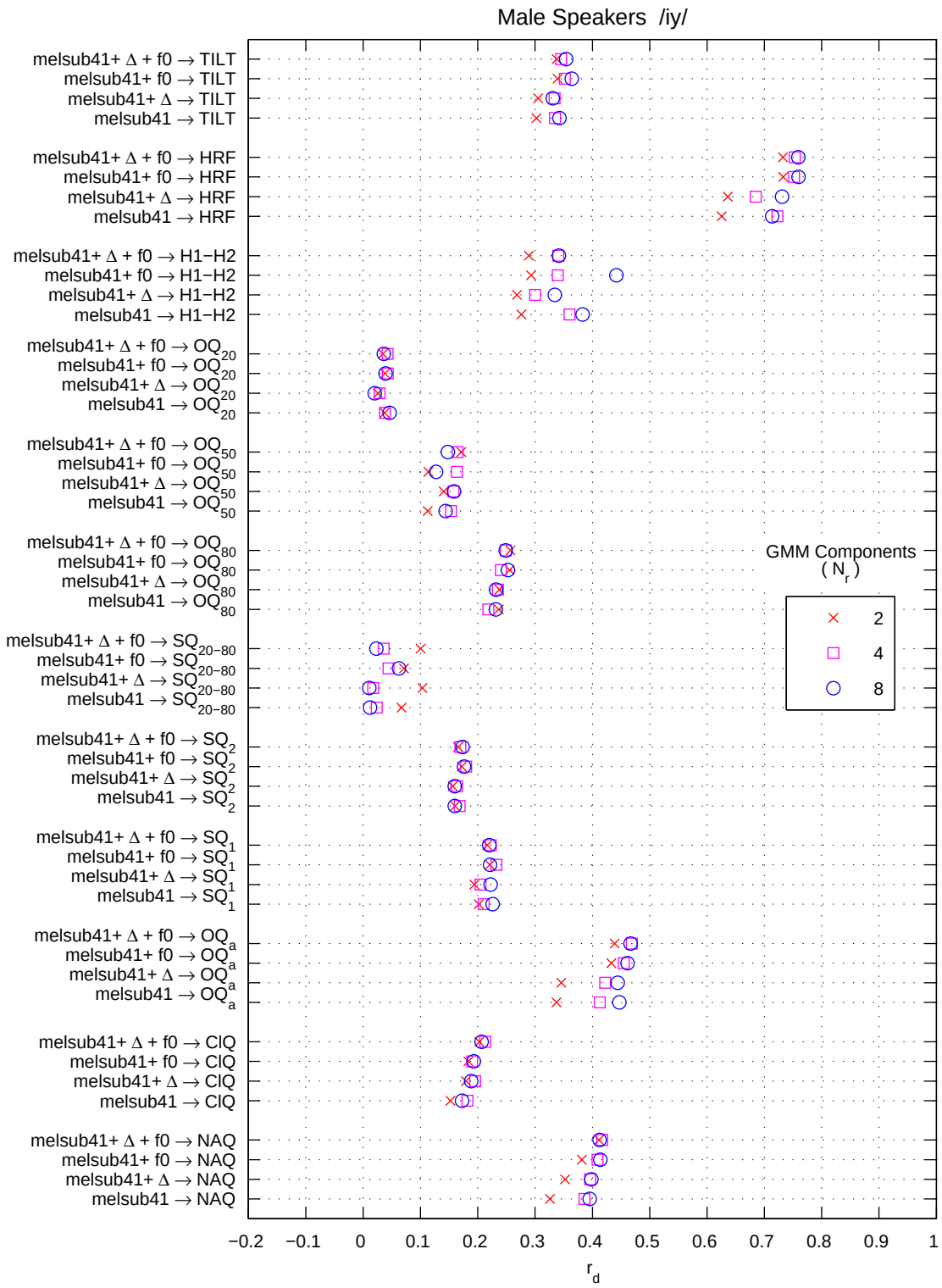


Figure 56: Coefficient of determination r_d between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /iy/, male speakers.

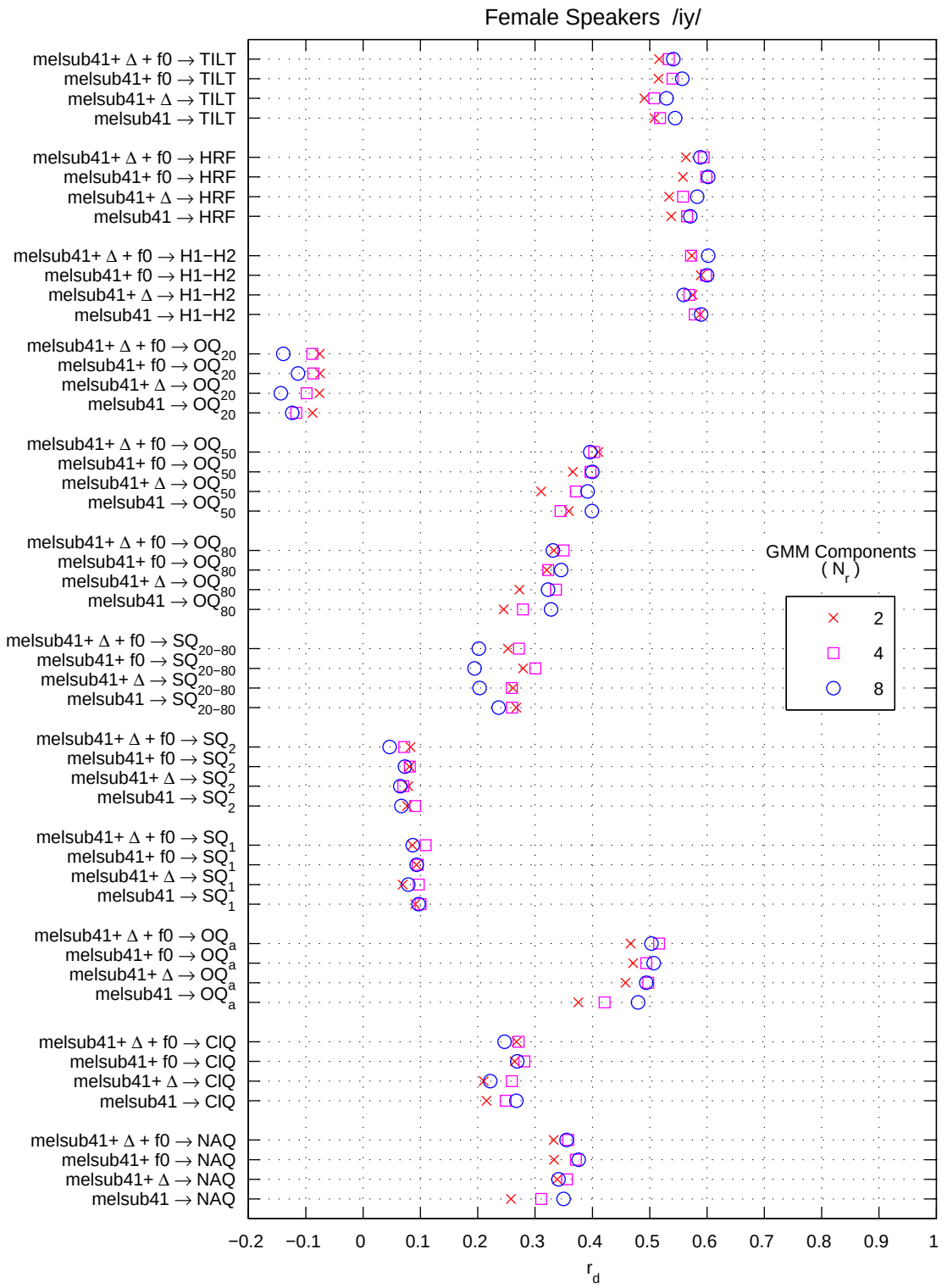


Figure 57: Coefficient of determination r_d between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /iy/, female speakers.

Figure 58: Coefficient of determination r_d between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ae/, male speakers.

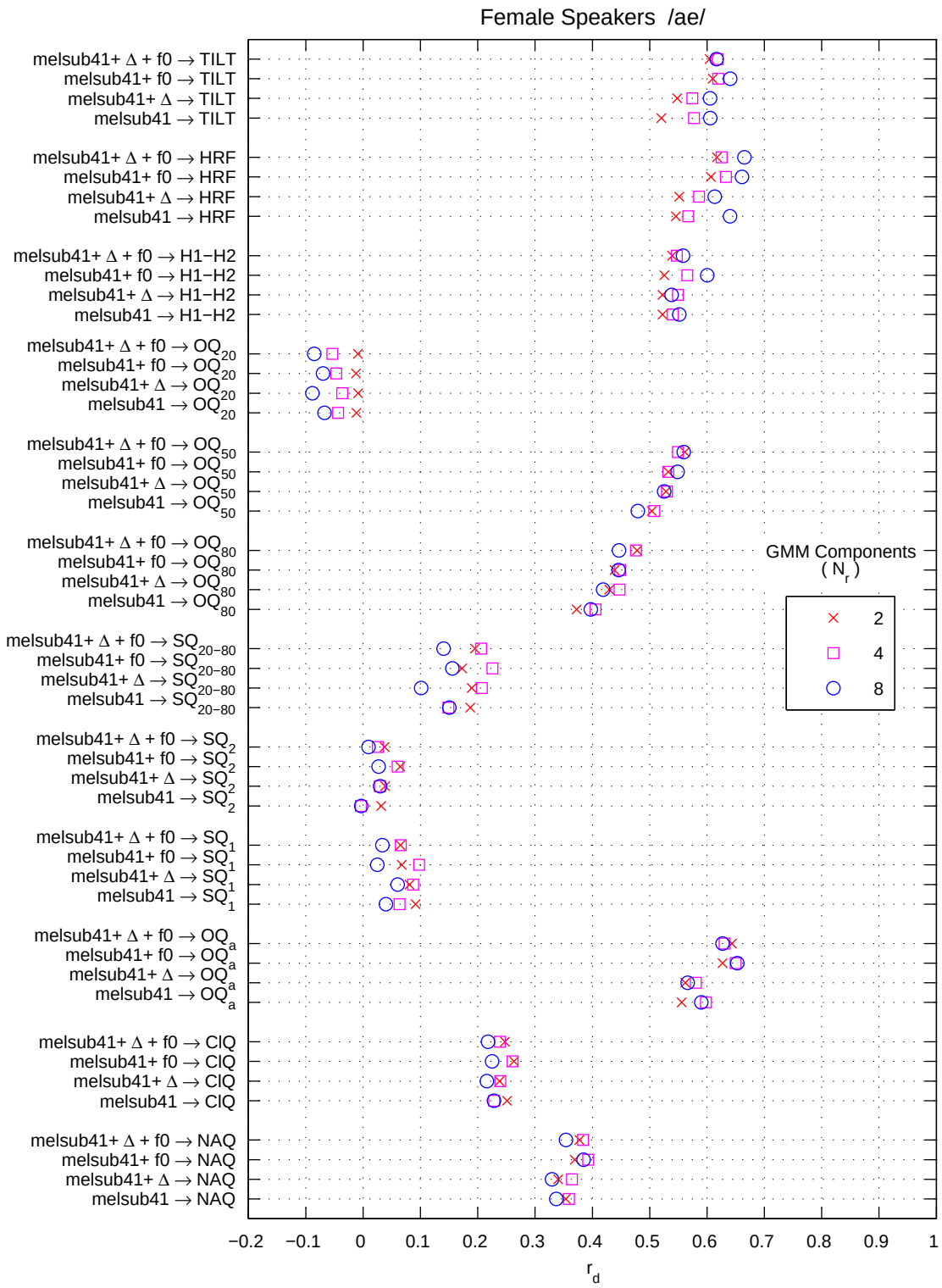


Figure 59: Coefficient of determination r_d between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ae/, female speakers.

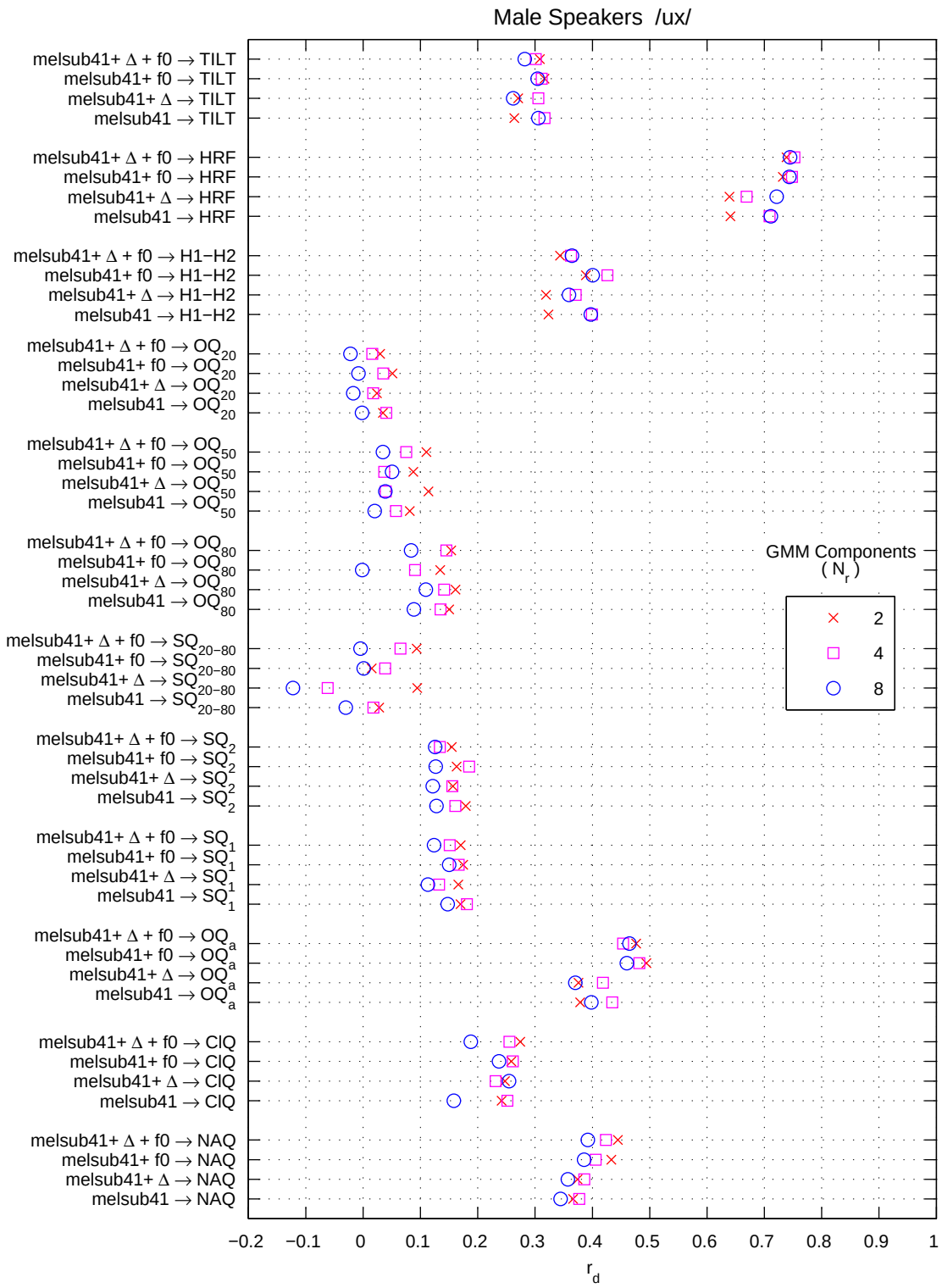


Figure 60: Coefficient of determination r_d between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ux/, male speakers.

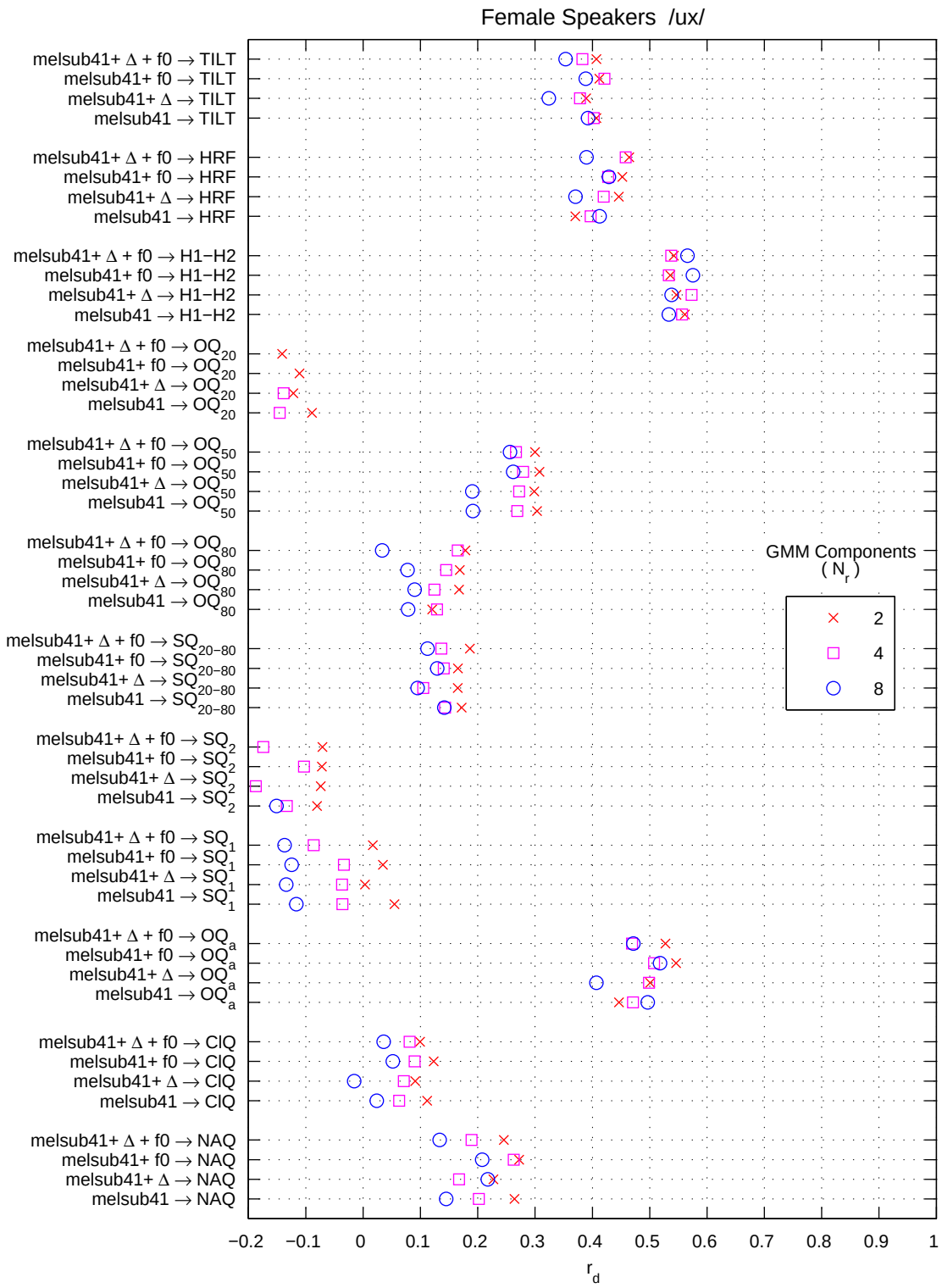


Figure 61: Coefficient of determination r_d between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). Direct measurement features, phoneme /ux/, female speakers.

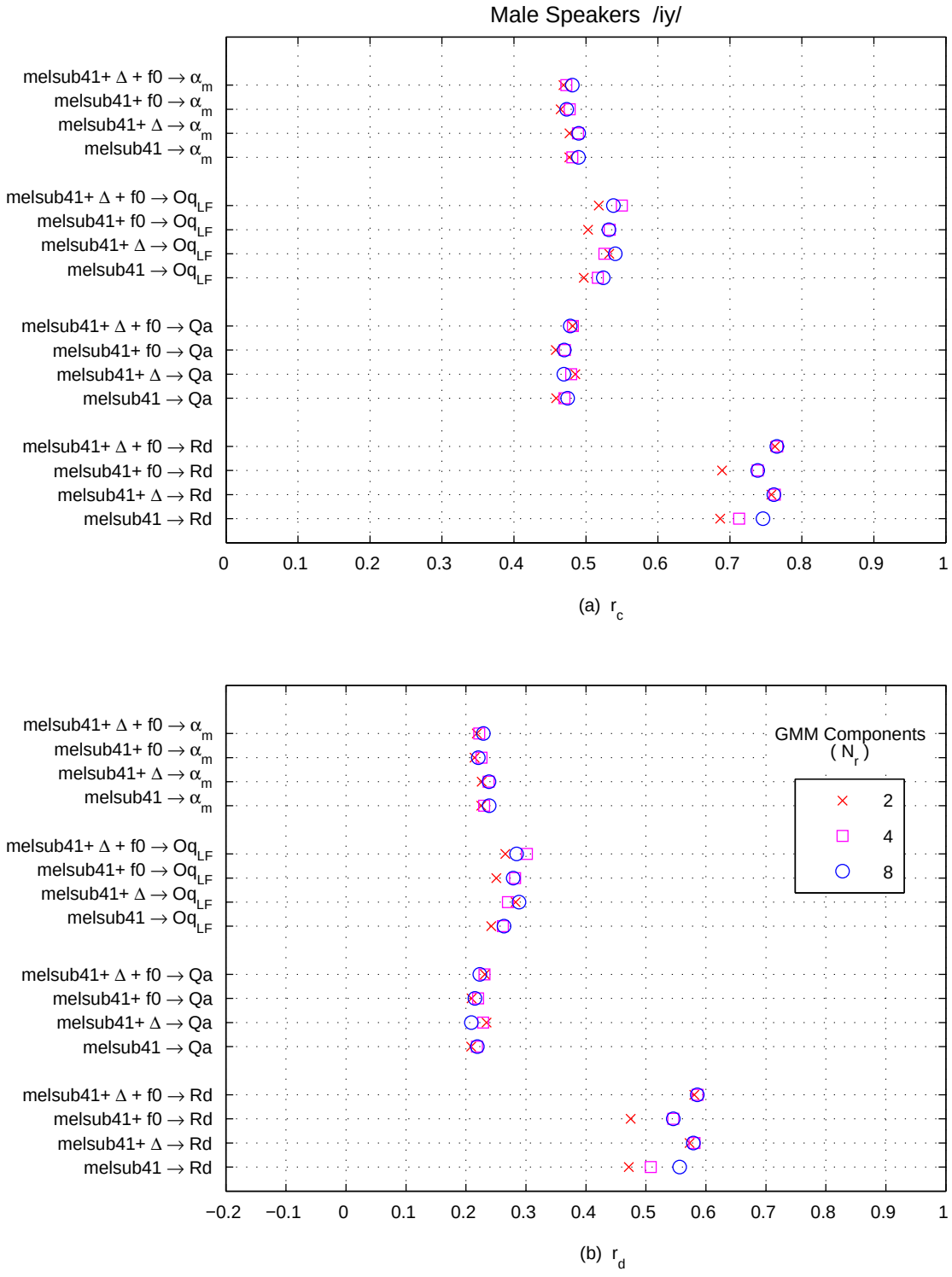


Figure 62: Correlation coefficient r_c (a) and coefficient of determination r_d (b) between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). LF-model features, phoneme /iy/, male speakers.

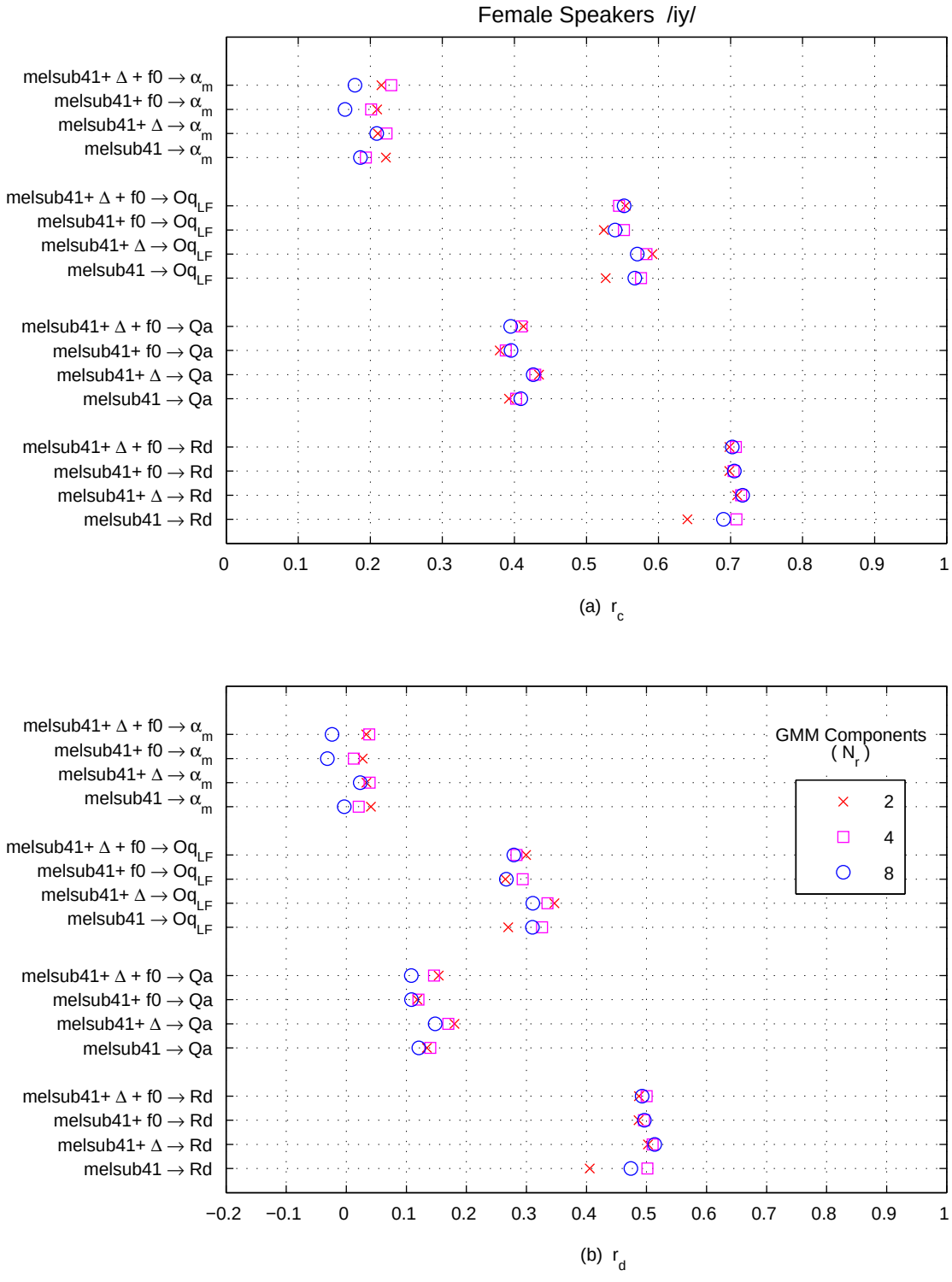


Figure 63: Correlation coefficient r_c (a) and coefficient of determination r_d (b) between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). LF-model features, phoneme /iy/, female speakers.

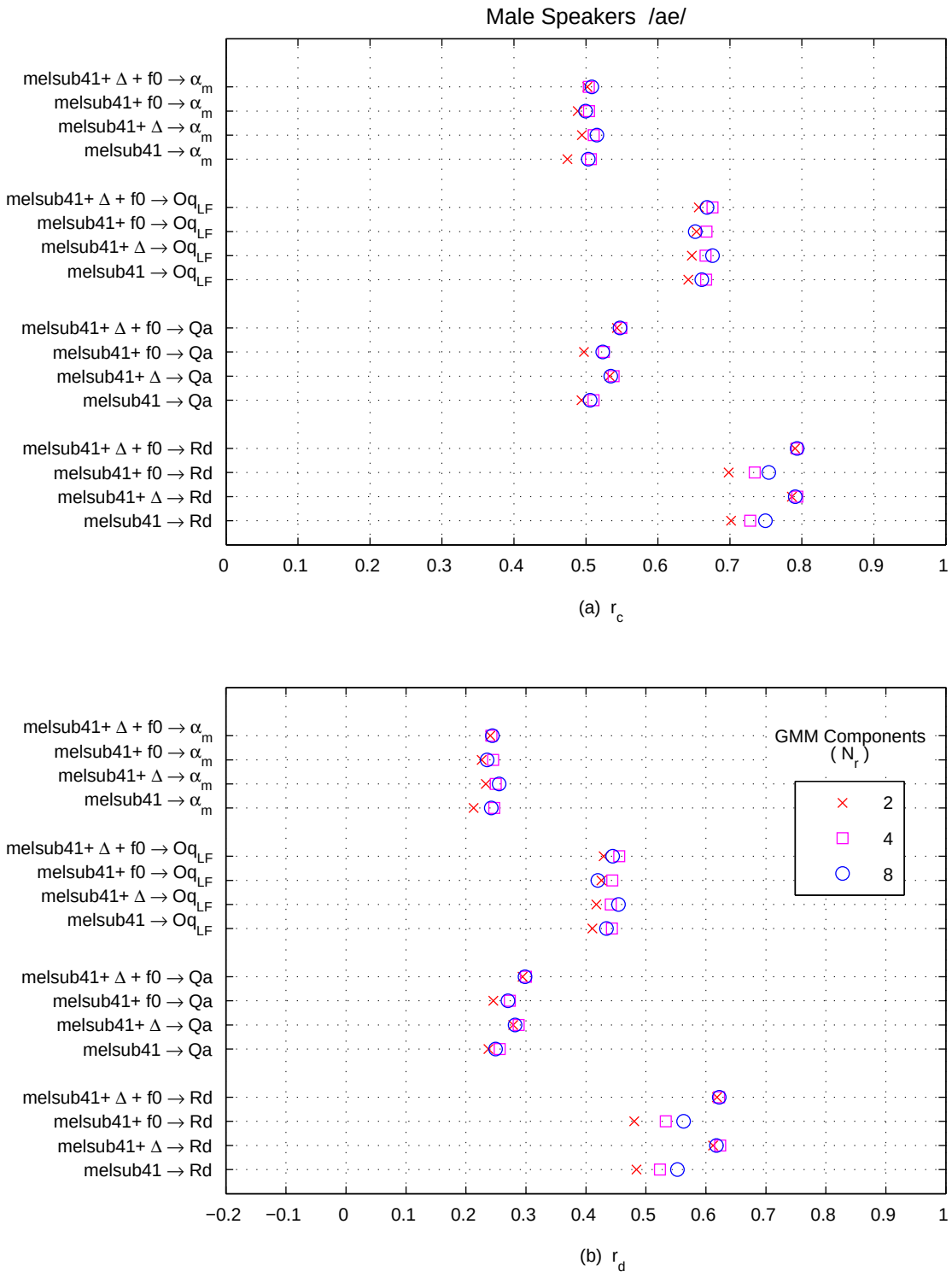


Figure 64: Correlation coefficient r_c (a) and coefficient of determination r_d (b) between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). LF-model features, phoneme /ae/, male speakers.

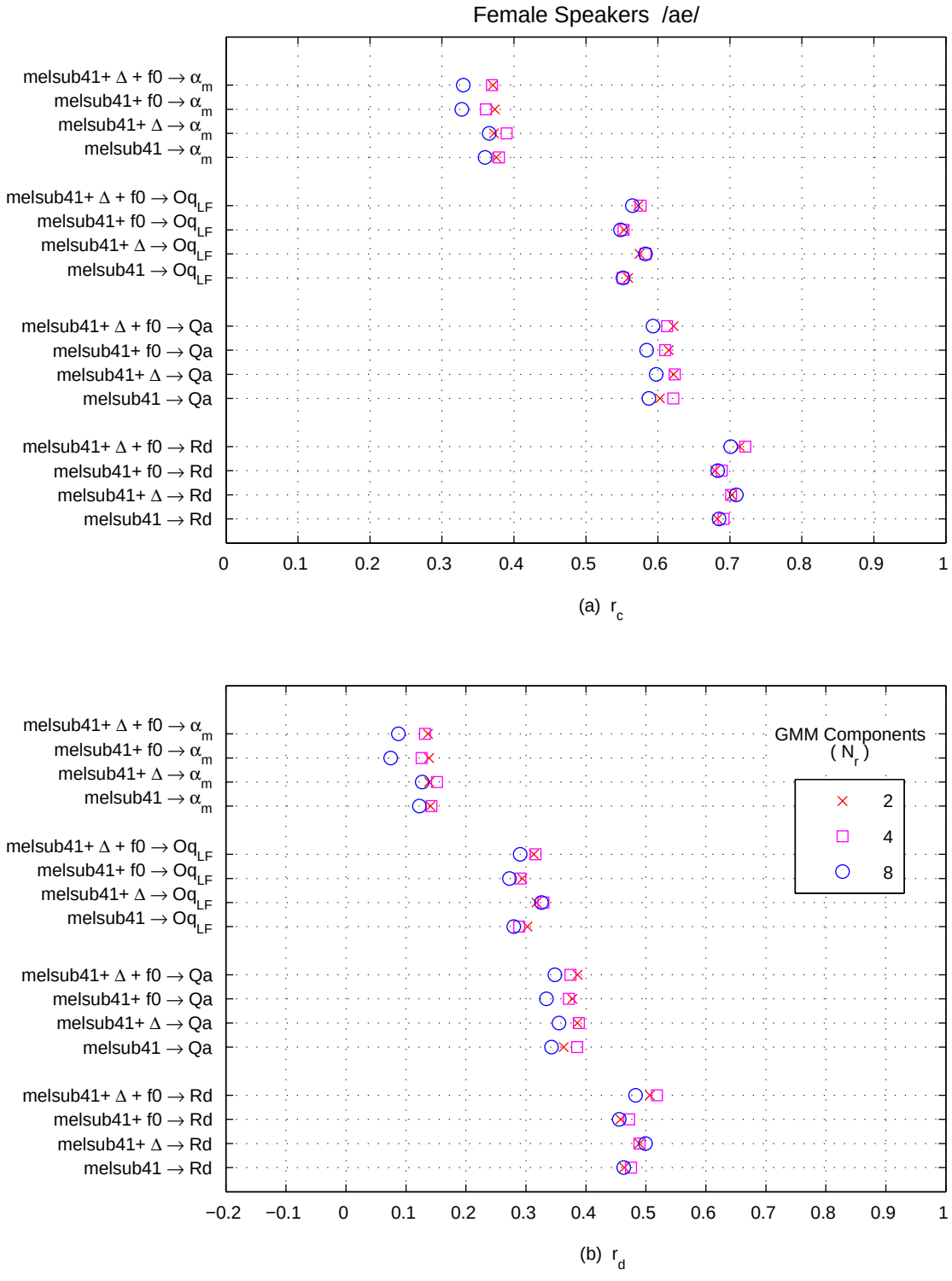


Figure 65: Correlation coefficient r_c (a) and coefficient of determination r_d (b) between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). LF-model features, phoneme /ae/, female speakers.

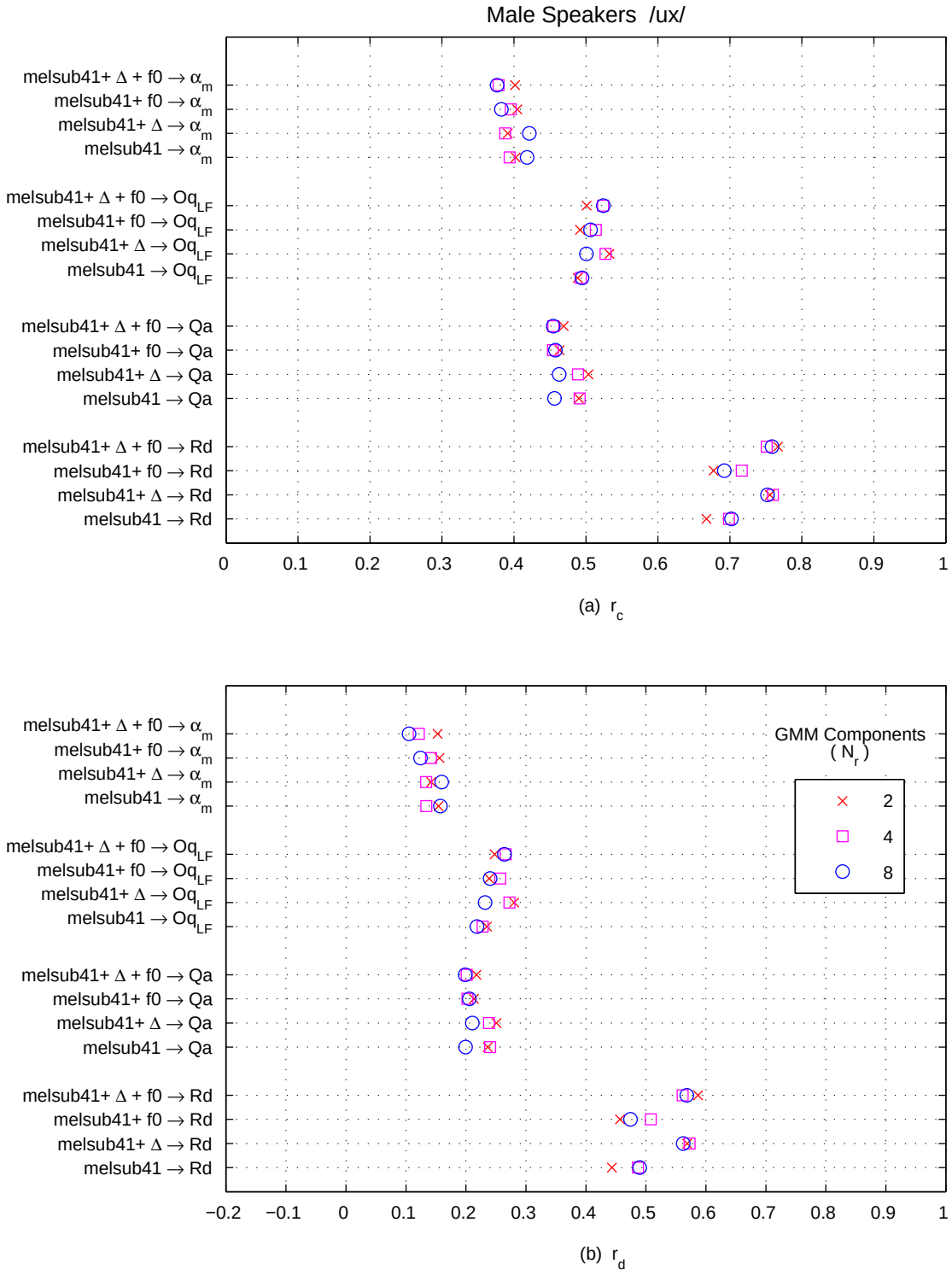


Figure 66: Correlation coefficient r_c (a) and coefficient of determination r_d (b) between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). LF-model features, phoneme /ux/, male speakers.

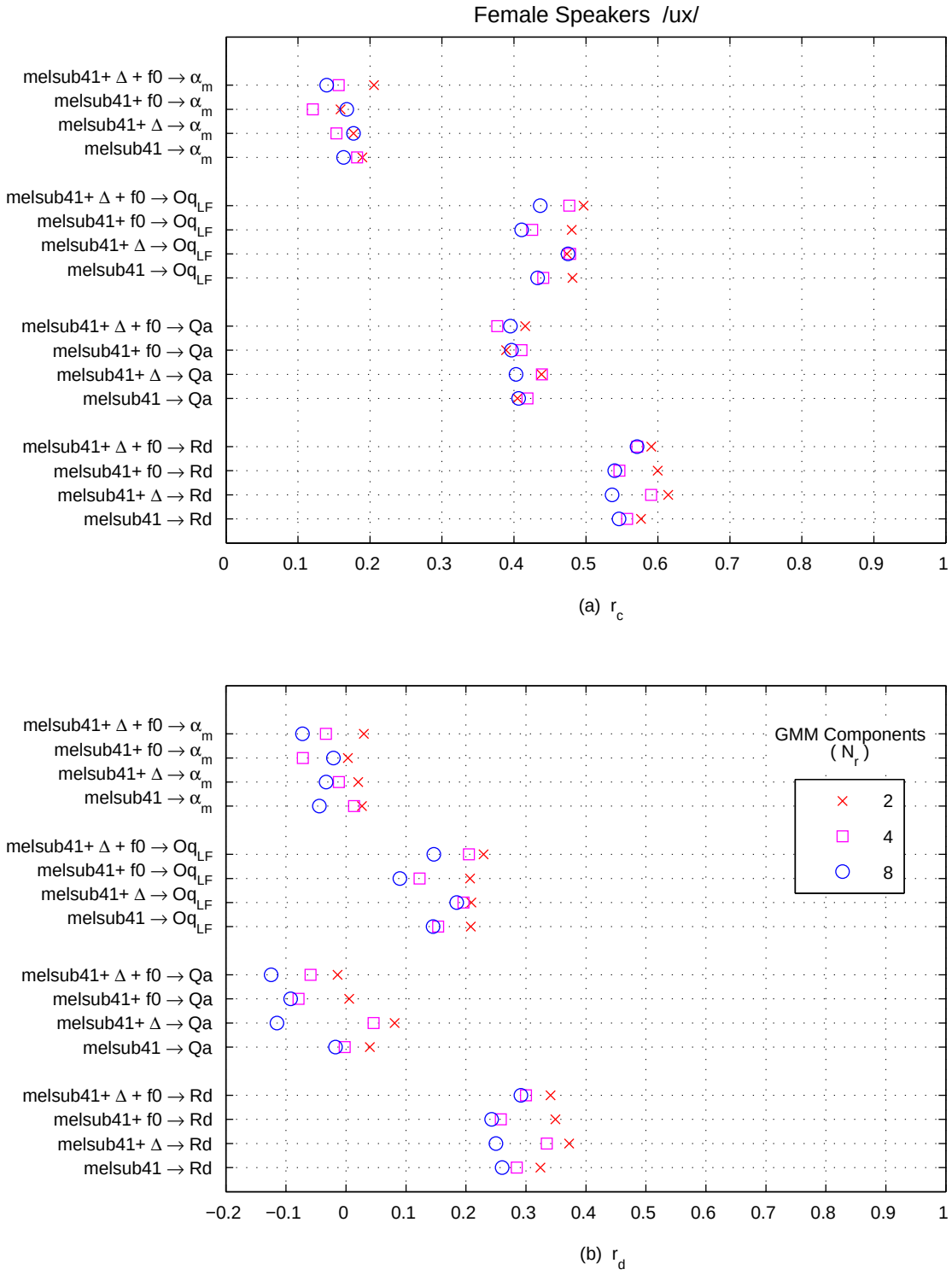


Figure 67: Correlation coefficient r_c (a) and coefficient of determination r_d (b) between IF and GMR glottal features. Effect of delta features (Δ) and pitch (f_0). LF-model features, phoneme /ux/, female speakers.

A.3 Joint Feature / Joint Phoneme GMMs

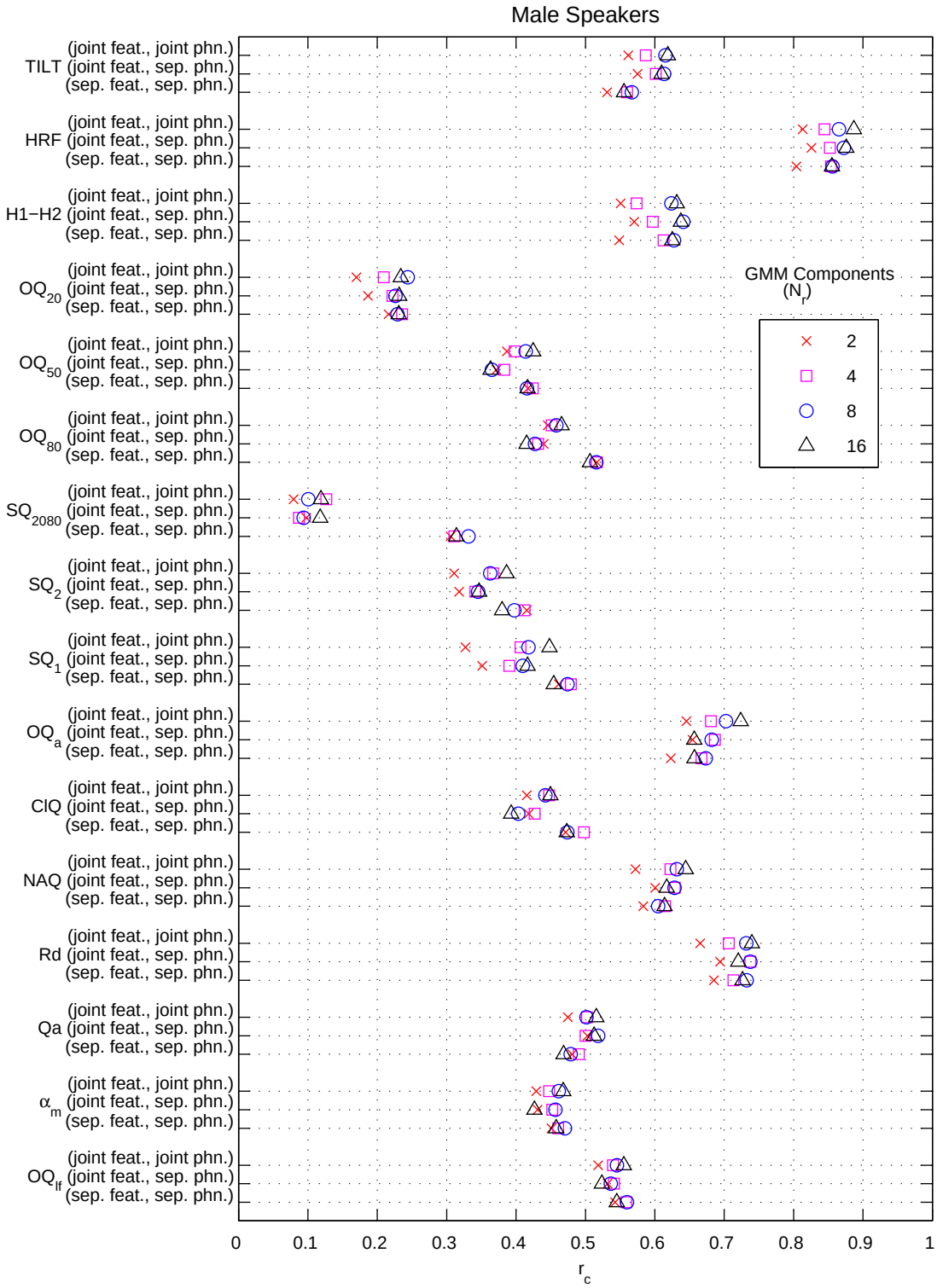


Figure 68: Correlation coefficient r_c between IF and GMR glottal features, effect of joint feature / joint phoneme GMMs compared against baseline, where a separate GMM was trained for each feature and phoneme. Male speakers.

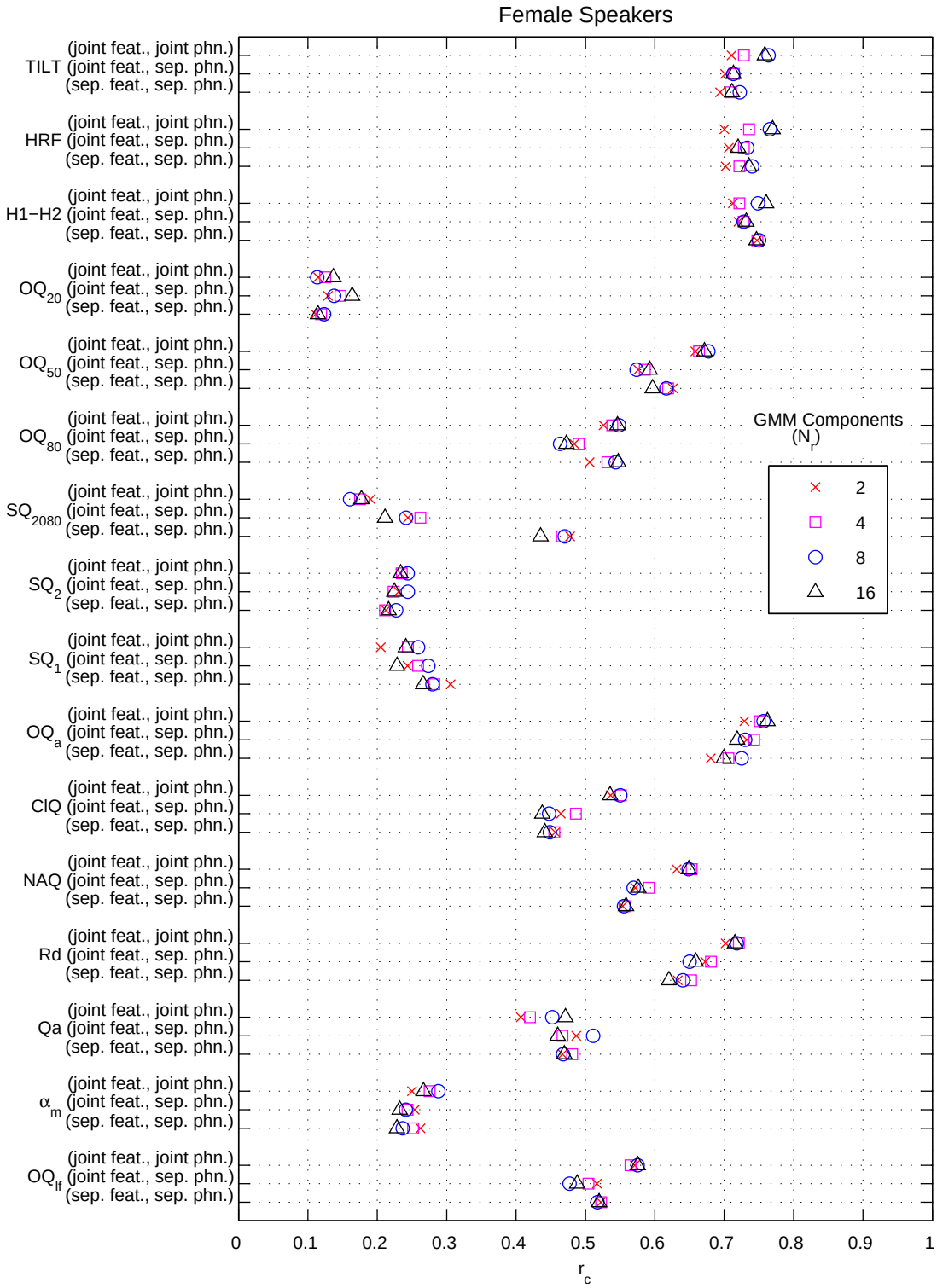


Figure 69: Correlation coefficient r_c between IF and GMR glottal features, effect of joint feature / joint phoneme GMMs compared against baseline, where a separate GMM was trained for each feature and phoneme. Female speakers.

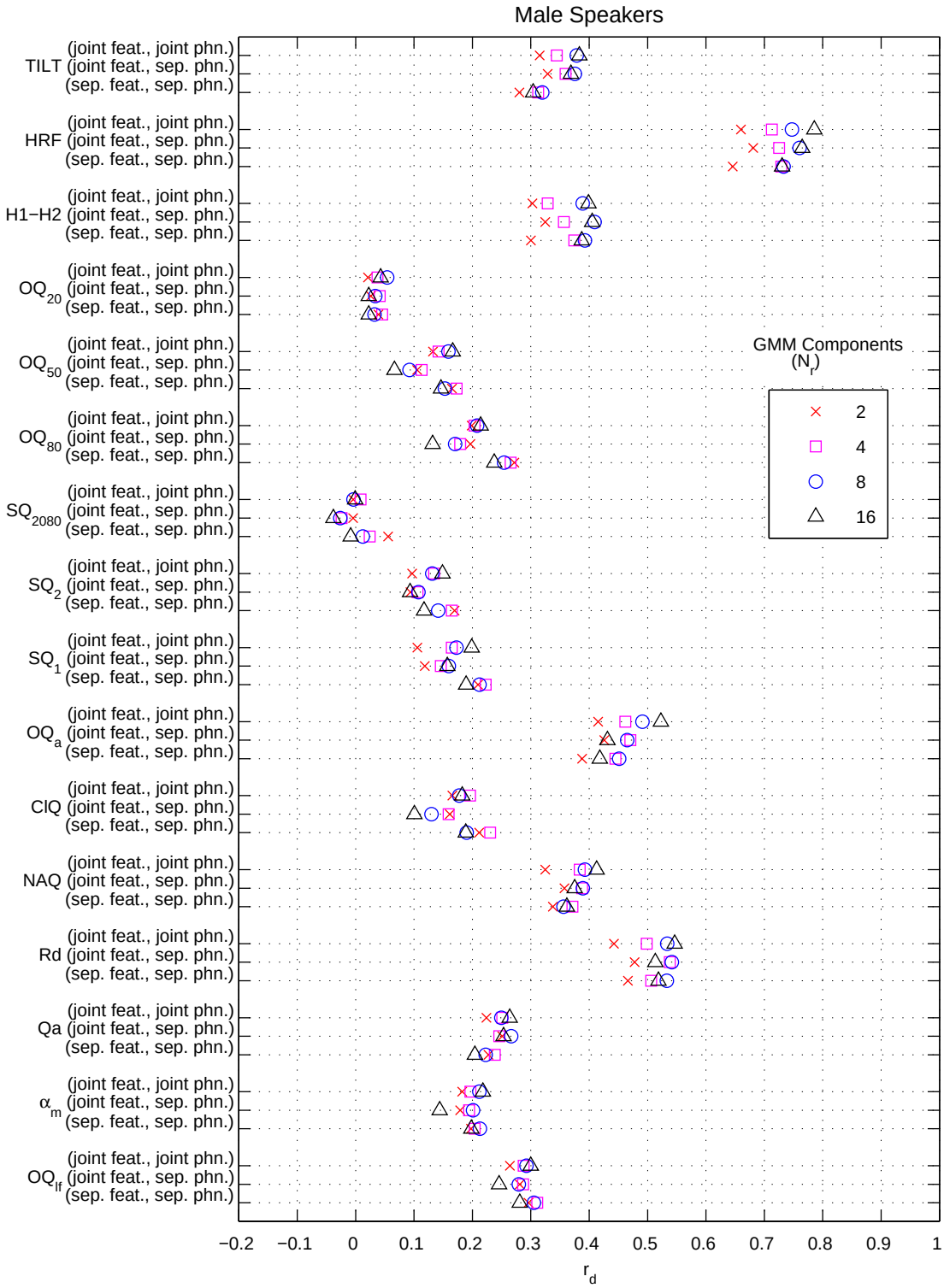


Figure 70: Coefficient of determination r_d between IF and GMR glottal features, effect of joint feature / joint phoneme GMMs compared against baseline, where a separate GMM was trained for each feature and phoneme. Male speakers.

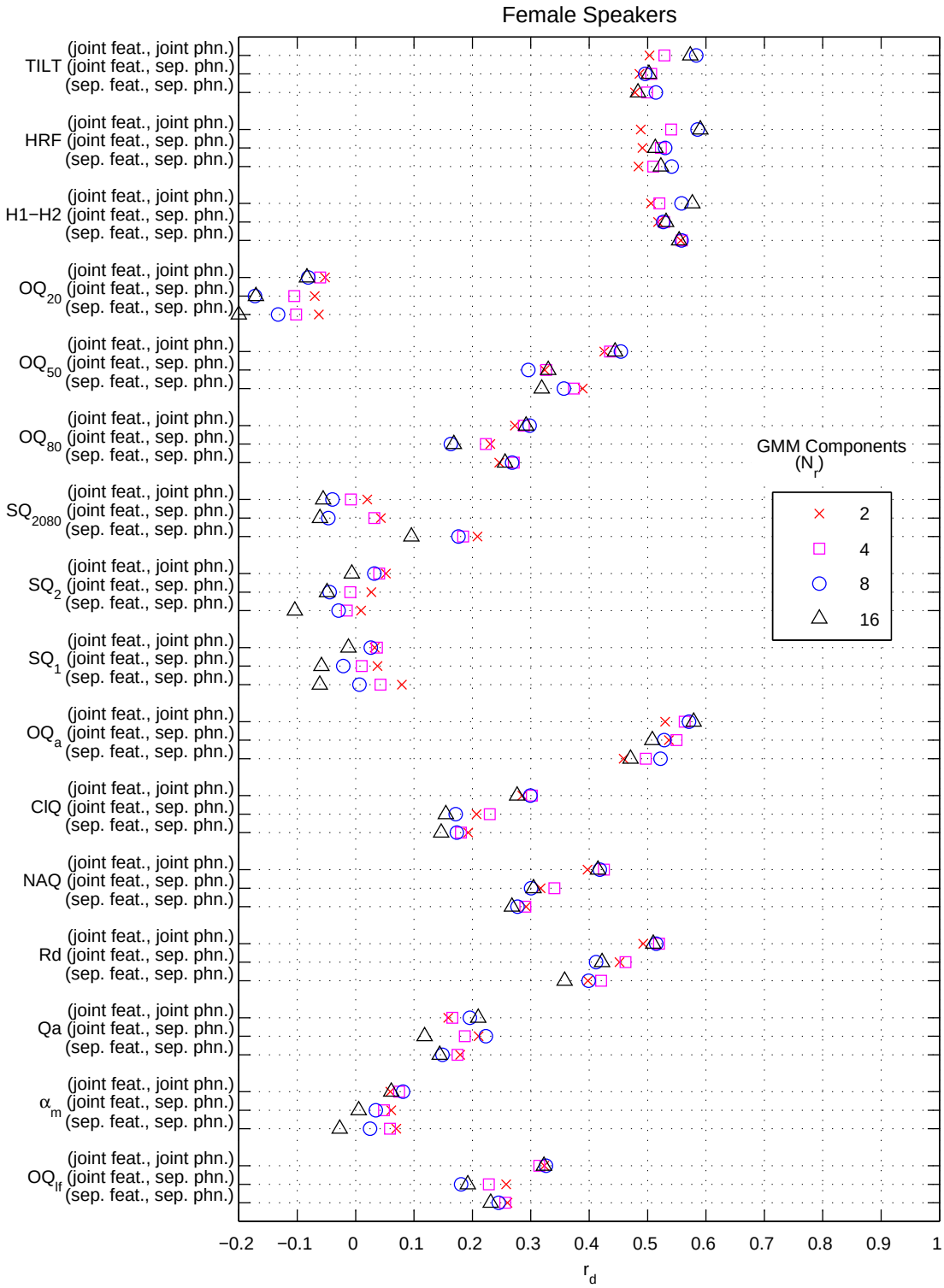


Figure 71: Coefficient of determination r_d between IF and GMR glottal features, effect of joint feature / joint phoneme GMMs compared against baseline, where a separate GMM was trained for each feature and phoneme. Female speakers.

A.4 Training on Additional Phonemes

Figure 72: Correlation coefficient r_c between IF and GMR glottal features, effect of training / testing on additional phonemes compared against baseline where only /iy/, /ae/, and /ux/ were processed. Male speakers.

Figure 73: Correlation coefficient r_c between IF and GMR glottal features, effect of training / testing on additional phonemes compared against baseline where only /iy/, /ae/, and /ux/ were processed. Female speakers.

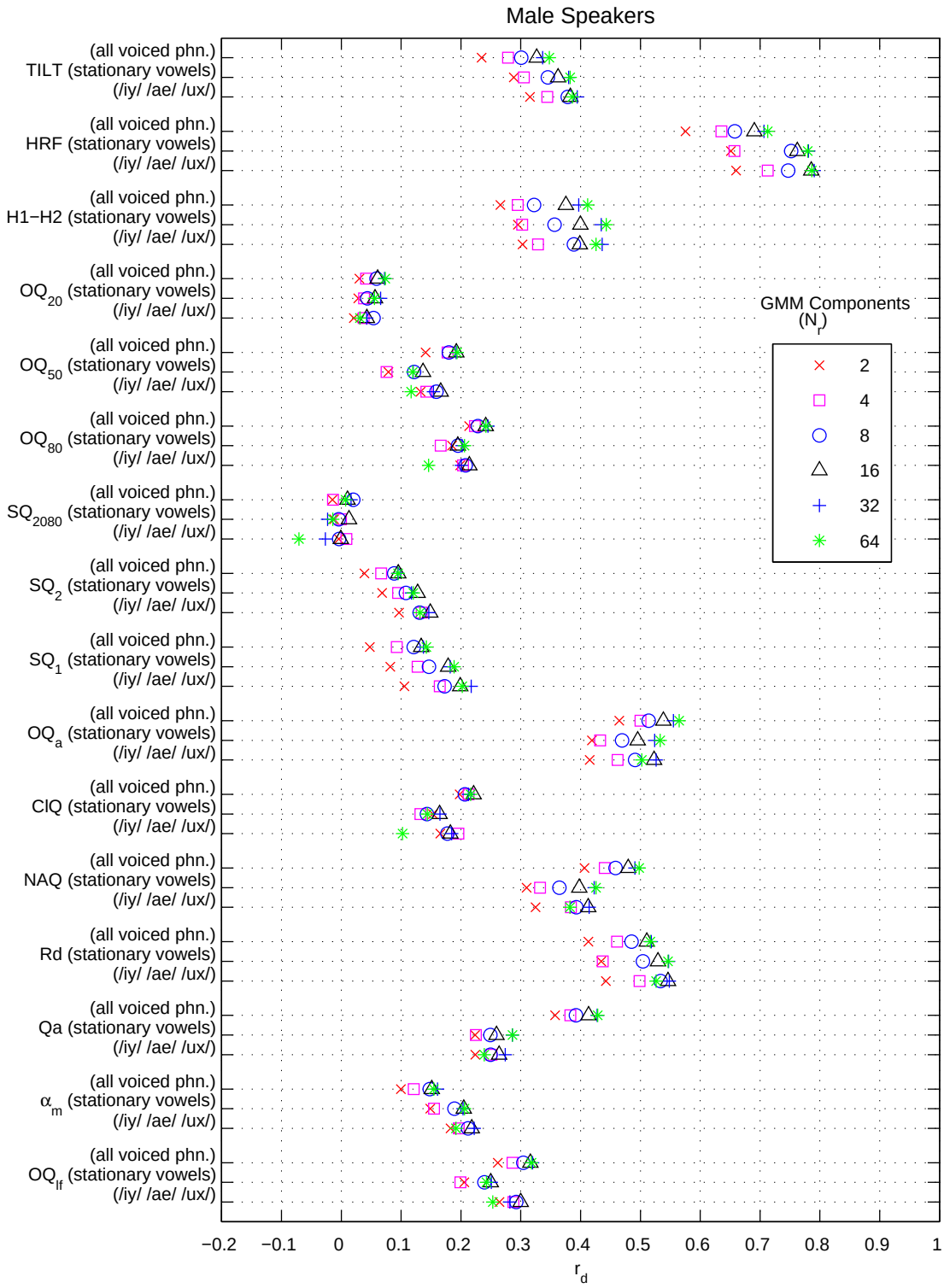


Figure 74: Coefficient of determination r_d between IF and GMR glottal features, effect of training / testing on additional phonemes compared against baseline where only /iy/, /ae/, and /ux/ were processed. Male speakers.

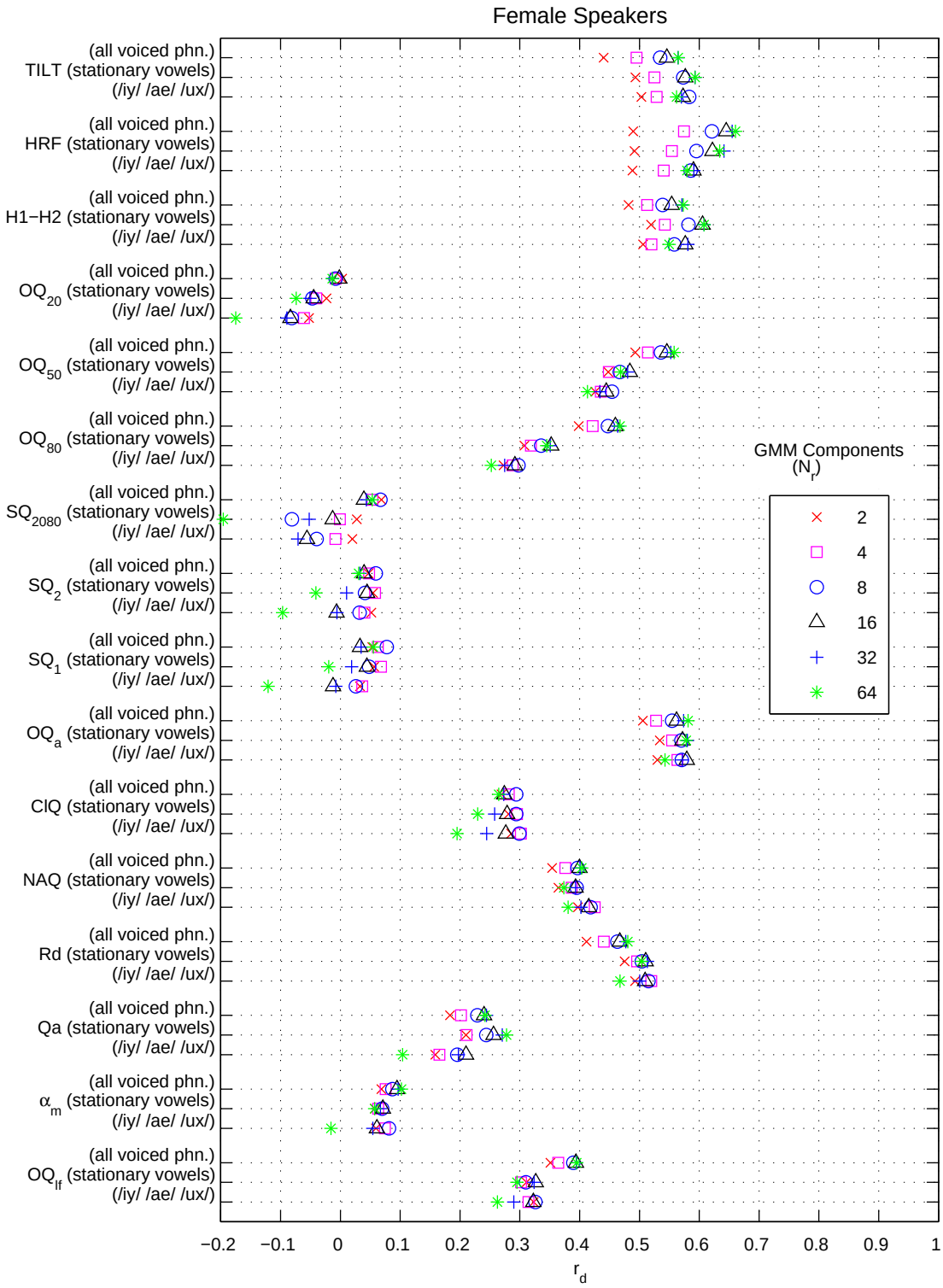


Figure 75: Coefficient of determination r_d between IF and GMR glottal features, effect of training / testing on additional phonemes compared against baseline where only /iy/, /ae/, and /ux/ were processed. Female speakers.

APPENDIX B

CORRESPONDENCE BETWEEN ELECTROGLOTTOGRAPH AND GLOTTAL WAVEFORM FEATURES

B.1 Introduction

The electroglottograph (EGG) signal can be regarded as a potential source of glottal features for the system proposed in Chapter 7, as it contains voice source information that is unaffected by the vocal tract and can be obtained with minimal inconvenience to the subject (Section 3.1.2). As the electroglottograph signal is a measure of vocal fold *contact*, the utility of EGG-derived features as outputs of the proposed Gaussian mixture regression (GMR) feature transformation procedure is dependent on the existence of a relationship between EGG (contact) and glottal waveform (airflow) features. This is because changes in airflow are what ultimately produce variations in the acoustic speech signal and its spectral envelope. Although a strong motivation for focusing on IF glottal features as GMR estimation targets on the remainder of this thesis stems from the unavailability of a sufficiently large EGG corpus of continuous speech,[1] initial experiments were performed to evaluate the correspondence between specific EGG and glottal airflow features. In what follows, these experiments are described and their results are discussed in an effort to evaluate the potential usefulness of the EGG signal within the context of the present work.

[1]The two EGG corpora initially considered for the research presented in this thesis contained either a large number of utterances (50) from a small number of speakers (7 males, 5 females) [67] or only 3 short utterances (and a total of 3 sentence texts) for 25 male and 27 female speakers [23]. Neither of these corpora contained the simultaneous phonetic and speaker diversity needed to train a generalizable, speaker-independent GMR model.)

B.2 EGG Features

Because the EGG signal transduces vocal fold contact, its salient features are somewhat different from those of the glottal waveform. Specifically, the information provided by the EGG waveform $l[n]$ is mostly related to the *opening* and *closing* phases of the glottal cycle. This is because during the open glottal phase, where the vocal folds are almost fully decontacted, the EGG signal becomes largely insensitive to further opening of the glottis and corresponding increases in glottal airflow. Conversely, during the closed phase, the EGG may register additional contact area between the vocal folds, due to their three-dimensional nature, even after the glottis has fully closed and no changes in airflow are taking place [15].

Generally, the glottal closure instant (GCI) is easily observable as a large positive peak in the derivative of the EGG signal $l'[n]$, as shown in Figure 76. Detection of the glottal opening instant, however, can be somewhat ambiguous due to strands of mucus that form across the glottis during the beginning of the opening phase [24], resulting in reduced impedance across the larynx. This is a possible explanation for the multiple negative peaks observed in Figure 76 in the vicinity of the opening phase. Such additional peaks may be due to the breaking of a *mucus bridge*, which causes a quick increase in impedance that may be misinterpreted on the EGG waveform as an instant of glottal opening even though the actual parting of the vocal folds (and the start of airflow) will have commenced sometime before.

In consideration of the EGG signal's limitations, the five time-domain EGG features used in this study focus on the contacting and decontacting phases of the EGG cycle. Four of these measures are intended to provide similar information to the airflow-based open quotient. The OQ_{eggXX} measures were obtained by thresholding [99, 101] the EGG signal at 20%, 50%, or 80% of its maximum amplitude, as shown in Figure 76. The interval of the glottal cycle above the threshold (high contact) was denoted as the closed glottal phase and the interval below the threshold (low contact)

as the open glottal phase. The open quotient measures were then computed as

$$OQ_{eggXX} = \frac{t_{cXX} - t_{oXX}}{T_0}, \tag{45}$$

where T_0 is the length of the pitch period, defined as the interval between consecutive EGG-derivative peaks t_{pk}. An additional OQ measure, OQ_{EGGpk}, was computed using Howard's method [64], where the positive peak of $l'[n]$ (t_{pk} in Figure 76) denotes the beginning of the closed glottal phase. Under this method, the start of the glottal open phase is defined as the instant where the EGG waveform crosses a 35% threshold of its maximum amplitude.

Because the EGG signal is insensitive during much of the open phase, the minima of $l[n]$ cannot be used to detect the point of maximum glottal flow. Therefore, an alternative definition of speed quotient, proposed in [101], was used. The EGG-based speed quotient was computed in a similar way to the glottal airflow feature $SQ_{20\text{-}80}$ (Section 6.2.2.1). Amplitude thresholds were set at 20% and 80% of the maximum EGG amplitude, and the positive and negative threshold crossing instants (t_{c20}, t_{c80}, t_{o80}, t_{o20}) were obtained as shown in Figure 76. The speed quotient was then computed as

$$SQ_{EGG} = \frac{t_{o20} - t_{o80}}{t_{c80} - t_{c20}}. \tag{46}$$

In an effort to explore possible relationships between other types of variation in the EGG waveform's shape and existing airflow features, the coarse time-domain structure of the EGG signal was represented as a set of DCT coefficients. For the k^{th} pitch cycle $l_k[n]$, which begins at the k^{th} GCI and ends at the $(k+1)^{th}$ GCI, a Hamming window $w[n]$ was applied to reduce edge effects. Then, to reduce the effect of pitch variations, $l_k[n]w[n]$ was resampled to 200 points, and the DCT was computed from the resampled, windowed EGG pitch cycle $\hat{l}_k[n]$ as follows:

$$D_m = \sqrt{\tfrac{2}{N}} \sum_{n=0}^{N-1} \hat{l}_k[n] \cos\left[\tfrac{\pi m}{N}(n+0.5)\right] \quad m = 0, \ldots, M-1. \tag{47}$$

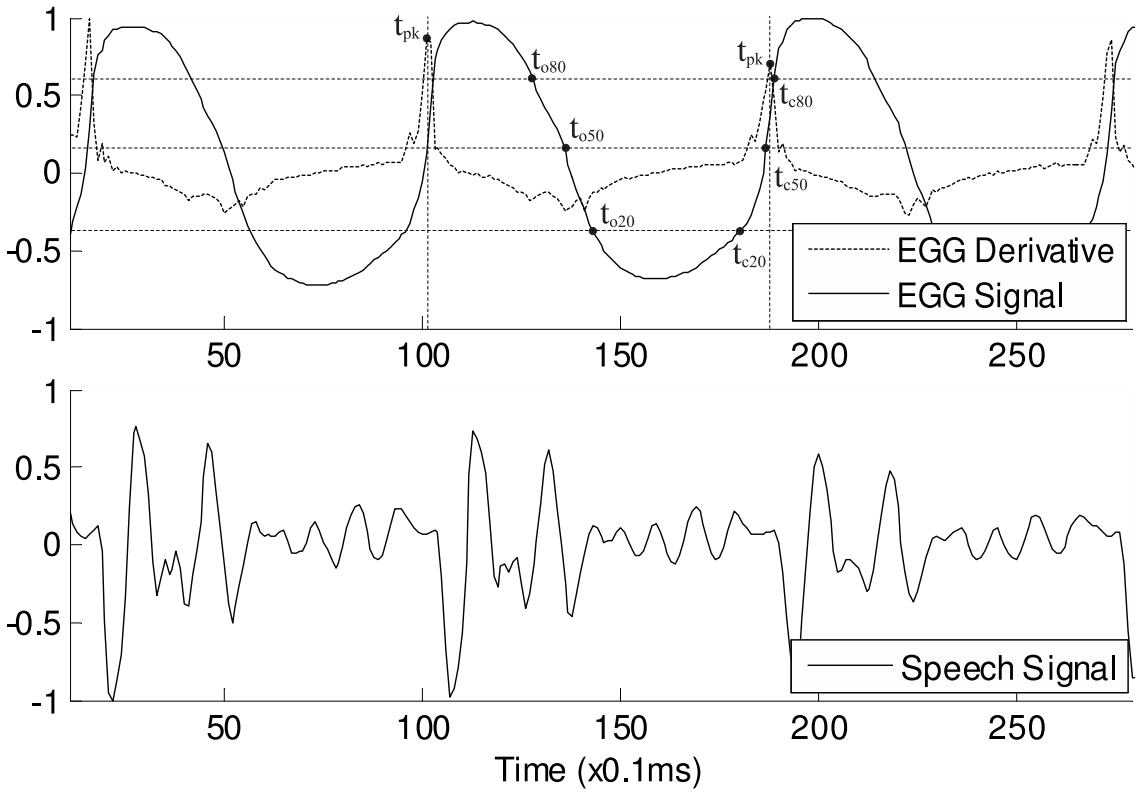


Figure 76: Electroglottograph (EGG) and acoustic speech signals for a vowel utterance.

The first 5 DCT coefficients $[D_1, \ldots, D_5]$ were retained, as they were found to contain, on average, over 98% of the energy in $\hat{l}[n]$. D_0 was not used, as it is a measure of DC-offset.

B.3 Speech Data and Feature Extraction

The data used for the experiments that follow came from the corpus by D.G. Childers, which may be obtained from the CD-ROM accompanying his book [23]. The corpus contains speakers with non-pathological larynges performing sustained utterances of 12 American English vowels. For each utterance, simultaneous acoustic and electroglottograph signals were recorded to digital media at a 10 kHz sampling rate (16-bit resolution). The acoustic speech signals were obtained using an Electro-Voice RE-10 microphone located 6 inches from the speaker's lips, and an electroglottograph by Synchrovoice, Inc. was used to obtain the EGG signals. A microphone correction transfer function was used to correct for low-frequency microphone distortions, as discussed in [25].

A set of 24 male and 24 female utterances of the vowel /AA/ (as in Bach) was selected for this study, as the high first-formant of the /AA/ vowel facilitates inverse filtering. Each utterance was roughly 2 seconds long. Because the EGG signal is measured near the larynx, the glottal events registered by the laryngograph occur sometime before the corresponding events in the acoustic speech signal, due to the physical delay caused by the length of the vocal tract and the distance between the subject and the microphone. (This effect can be observed in Figure 76, where the glottal closure instant, as evidenced by the positive peak of the EGG derivative, still occurs slightly before the corresponding energy increase in the speech signal despite an approximate correction for the time-delay.) The exact time delay is dependent on the vocal tract length of each speaker, and must be accurately determined to allow for simultaneous, pitch-synchronous analysis of the EGG and speech signals. The

speaker-dependent delay between the EGG signal $l[n]$ and the acoustic speech signal $s[n]$ was determined by obtaining estimates of the glottal closure instants (GCI) from $s[n]$ via the DYPSA algorithm [85], and computing their average distance to the locations of the positive peaks t_{pk} of the EGG derivative $l'[n]$ (i.e., the EGG-based GCI estimates). The delay d_l between $l[n]$ and $s[n]$ was found to be typically between 9 to 11 samples. Lower-frequency variations in the EGG signal, which occur due to movement of the larynx and other neck structures [15, 27], and are not related to the vocal fold vibratory pattern, were removed using a zero-phase high-pass filter with a 30 Hz cutoff.

Sustained vowel utterances are particularly amenable to inverse filtering (IF) because an approximately stationary vocal tract can be assumed. Glottal waveform estimates were obtained via multiple-window closed-phase inverse filtering, in which data from the closed glottal phase of each pitch period are combined to derive a more robust estimate of the vocal tract filter (VTF) via the inclusion of additional normal equations [18]. The length of each closed phase was set to $2p + 1$, where p, the VTF order, was manually adjusted for each utterance to obtain glottal waveform estimates with minimum ripple (by visual inspection). The optimal values of p varied from 10 to 14. Slight manual adjustments the EGG delay d_l were also made as to ensure that the EGG-estimated GCIs coincided with the onset of glottal closure on the glottal waveform estimates.

EGG and glottal waveform features were extracted pitch-synchronously from each utterance, with one observation per pitch cycle. Outlier removal was performed as described in Section 6.3.1. This resulted in 9,477 and 5,358 observations for female and male speakers, respectively. The mean and standard deviation for each IF and EGG feature are shown in Tables 24 and 25, respectively. From these tables, an important difference between EGG and IF features can already be observed. While all of the open quotient (OQ) features obtained from IF features show largely different

Table 24: Mean value and standard deviation for each EGG feature, obtained from 24 male and 24 female sustained utterances of the vowel /AA/.

	Males		Females	
	Mean	Std.	Mean	Std.
OQ_{EGG20}	0.423	0.054	0.421	0.084
OQ_{EGG50}	0.560	0.053	0.585	0.072
OQ_{EGG80}	0.712	0.041	0.753	0.059
OQ_{EGGpk}	0.540	0.057	0.551	0.077
SQ_{EGG}	5.72	3.3	4.80	1.8

means across genders, the EGG-based OQ features do not vary much between males and females. This result suggests a similar conclusion as the studies in [37, 101], where airflow-based OQ was found to vary significantly across different vocal intensities while EGG-based OQ did not. Because the open quotient is known to vary with respect to vocal intensity, the authors concluded that the EGG-based "OQ" measures are not indicative of the relative duration of the open and closed phases of the glottal airflow cycle.

B.4 Rank Correlation between Glottal Waveform and EGG Time-Domain Features

An initial assessment of the relationship between glottal airflow and vocal fold contact features was obtained by examining the Spearman rank-correlation coefficient r_r (defined in Equation 9) between each IF and EGG feature pair. All experiments were conducted separately for each gender. As discussed in Section 4.4, r_r can be used to measure the strength of a possibly non-linear, monotonic relationship between a pair of features, with $r_r \approx 1$ indicating that one feature can be well-approximated as a monotonic function of the other. The r_r values across all speakers, shown in Table 26, indicate relationships between feature pairs that are usually weak, with nearly all values of $|r_r|$ below 0.5. The speaker-dependent nature of the correlations between EGG and IF features is evidenced in Table 27, which shows the 25^{th} and

Table 25: Mean value and standard deviation for each glottal waveform feature, obtained from 24 male and 24 female sustained utterances of the vowel /AA/ via inverse filtering (IF).

	Males		Females	
	Mean	Std.	Mean	Std.
OQ_{LF}	0.439	0.15	0.781	0.12
α_m	0.826	0.015	0.838	0.022
Qa	0.140	0.075	0.310	0.16
Rd	1.20	0.49	2.24	0.69
NAQ	0.163	0.068	0.183	0.041
ClQ	0.381	0.15	0.361	0.083
OQa	0.395	0.14	0.761	0.13
SQ_1	1.06	0.49	1.85	0.61
SQ_2	0.518	0.31	1.45	0.62
$SQ_{20\text{-}80}$	0.794	0.73	2.51	1.1
OQ_{20}	0.478	0.16	0.760	0.13
OQ_{50}	0.231	0.081	0.542	0.11
OQ_{80}	0.100	0.050	0.316	0.088
$H1\text{-}H2$	-5.97	0.82	-6.11	0.86
HRF	15.7	2.9	11.4	1.5
$TILT$	-37.0	4.6	-47.2	5.9

75^{th} percentiles of r_r when correlation is computed separately on individual speakers. The results show that there is moderate-to-high correlation ($0.60 > |r_r| > 0.76$) on a few feature pairs for more than one speaker. However, the low $|r_r|$ values at the opposite percentile indicate that these relationships do not persist across speakers. In some cases (e.g., $TILT$ for male speakers), the interquartile range shows a change in sign, indicating that the relationship between the EGG and IF feature tends to be monotonically increasing for some speakers, and monotonically decreasing for others.

B.5 Estimation of Airflow Features from Contact Features

While the rank-correlation coefficient is a useful tool for measuring relationships between pairs of features, it does not provide information about possible relationships between an IF feature and a *set* of EGG features. To investigate this possibility, the GMR feature transformation method was applied to the estimation of glottal waveform features (obtained via IF) through the transformation of a set of EGG features. For each gender and each IF feature, GMMs were trained using all 24 speakers, with the number of Gaussian mixtures N_r varying from 2 to 16. For each IF feature, the GMMs were trained using either the set of EGG DCT coefficients $\{D_1, \ldots, D_5\}$ or the set of five time-domain EGG features $\{OQ_{EGG20}, OQ_{EGG50}, OQ_{EGG80}, OQ_{EGGpk}, SQ_{EGG}\}$. Two-thirds of the observations were randomly selected for training and the rest were used for testing. In this way, both the training and testing data contained observations from the same speakers, producing a best-case-scenario that is helpful in detecting any potential multivariate relationships.

The ability to transform a set of EGG features into each IF feature was evaluated by the linear correlation coefficient r_c between the IF feature observations and the estimates produced by transforming either the time-domain or DCT EGG features using the GMR procedure (Section 4.3). The correlation coefficient r_c was computed separately for each speaker. The 25^{th}, 50^{th}, and 75^{th} percentiles are given in Tables

Table 26: Spearman rank-correlation coefficient r_r between IF and EGG features obtained from 24 male and 24 female sustained utterances of the vowel /AA/.

		OQ_{EGG20}	OQ_{EGG50}	OQ_{EGG80}	OQ_{EGGpk}	SQ_{EGG}
OQ_{LF}	Males	0.43	0.40	0.31	0.36	-0.18
	Females	-0.16	-0.10	-0.11	-0.06	0.01
α_m	Males	0.22	0.28	0.23	0.28	-0.14
	Females	0.07	0.08	0.11	0.05	-0.11
Qa	Males	0.34	0.11	-0.04	0.11	-0.06
	Females	0.07	0.13	0.15	0.06	-0.19
Rd	Males	0.26	-0.02	-0.15	0.04	-0.06
	Females	0.04	0.16	0.21	0.09	-0.22
NAQ	Males	0.06	-0.17	-0.28	-0.13	-0.02
	Females	0.21	0.26	0.22	0.23	-0.37
ClQ	Males	-0.05	-0.32	-0.40	-0.25	-0.01
	Females	0.06	0.15	0.16	0.11	-0.24
OQa	Males	0.44	0.35	0.16	0.35	-0.39
	Females	0.06	0.09	0.07	0.03	0.10
SQ_1	Males	0.25	0.48	0.48	0.40	-0.06
	Females	-0.10	-0.14	-0.15	-0.12	0.16
SQ_2	Males	0.33	0.57	0.57	0.54	-0.33
	Females	-0.16	-0.25	-0.25	-0.25	0.21
$SQ_{20\text{-}80}$	Males	0.45	0.33	0.12	0.33	-0.04
	Females	-0.12	-0.14	-0.12	-0.10	0.10
OQ_{20}	Males	-0.22	-0.14	-0.08	-0.17	-0.06
	Females	-0.15	-0.05	-0.09	-0.01	-0.24
OQ_{50}	Males	0.22	0.10	0.00	0.14	-0.29
	Females	-0.36	-0.32	-0.37	-0.35	-0.08
OQ_{80}	Males	0.11	0.06	-0.03	0.10	-0.42
	Females	-0.40	-0.35	-0.39	-0.37	-0.14
$H1\text{-}H2$	Males	0.13	0.01	-0.02	-0.02	0.13
	Females	-0.05	-0.09	-0.11	-0.10	0.02
HRF	Males	-0.35	-0.20	-0.02	-0.24	0.32
	Females	-0.07	-0.18	-0.23	-0.11	0.22
$TILT$	Males	-0.32	-0.22	-0.12	-0.19	0.44
	Females	-0.05	-0.13	-0.15	-0.07	0.11

Table 27: Spearman rank-correlation coefficient r_r between IF and EGG features. Interquartile-range (IQR) across speakers.

		OQ_{EGG20}	OQ_{EGG50}	OQ_{EGG80}	OQ_{EGGpk}	SQ_{EGG}
OQ_{LF}	Males	-.02, .46	-.05, .58	-.12, .49	-.05, .61	-.40, .12
	Females	-.09, .12	-.17, .16	-.15, .10	-.17, .16	-.19, .24
α_m	Males	-.20, .16	-.26, .13	-.22, .11	-.24, .11	-.19, .25
	Females	-.10, .12	-.06, .19	-.06, .14	-.05, .23	-.13, .06
Qa	Males	-.21, .23	-.13, .25	-.06, .22	-.15, .22	-.33, -.04
	Females	-.08, .10	-.03, .22	-.05, .15	-.01, .22	-.15, .09
Rd	Males	-.17, .25	-.17, .24	-.05, .21	-.18, .22	-.29, .00
	Females	-.10, .06	-.09, .25	-.06, .19	-.04, .23	-.26, .10
NAQ	Males	-.21, .35	-.24, .34	-.21, .32	-.27, .28	-.27, .16
	Females	-.16, .52	-.17, .37	-.08, .41	-.18, .42	-.41, .12
ClQ	Males	-.20, .31	-.30, .22	-.25, .11	-.33, .15	-.15, .16
	Females	-.18, .19	-.16, .24	-.10, .20	-.16, .29	-.23, .17
OQa	Males	.07, .52	.13, .56	.03, .51	.10, .56	-.49, .15
	Females	-.13, .21	-.06, .29	-.13, .29	-.01, .22	-.15, .13
SQ_1	Males	-.19, .50	-.18, .54	-.21, .39	-.12, .57	-.27, .15
	Females	-.19, .18	-.23, .26	-.20, .14	-.26, .30	-.13, .29
SQ_2	Males	-.09, .33	-.08, .35	-.18, .28	-.09, .38	-.20, .12
	Females	-.10, .23	-.17, .16	-.16, .10	-.21, .17	-.13, .18
$SQ_{20\text{-}80}$	Males	.10, .60	.26, .73	.03, .58	.29, .76	-.41, -.09
	Females	-.30, .26	-.38, .39	-.25, .18	-.32, .42	-.19, .33
OQ_{20}	Males	-.30, .27	-.30, .20	-.25, .20	-.37, .20	-.27, .16
	Females	-.30, .26	-.22, .28	-.21, .21	-.21, .29	-.18, .42
OQ_{50}	Males	-.14, .49	.00, .28	-.14, .27	-.02, .29	-.50, .05
	Females	-.34, .28	-.15, .30	-.12, .25	-.24, .33	-.12, .44
OQ_{80}	Males	-.16, .56	-.15, .59	-.31, .45	-.07, .60	-.62, .02
	Females	-.28, .27	.05, .36	.02, .32	-.12, .41	-.24, .27
$H1\text{-}H2$	Males	-.11, .04	-.17, .03	-.14, .05	-.13, .03	-.06, .04
	Females	-.02, .06	-.07, .04	-.04, .02	-.05, .05	-.04, .05
HRF	Males	-.52, -.11	-.54, -.18	-.45, -.11	-.61, -.15	.10, .51
	Females	-.20, .16	-.29, .07	-.27, .01	-.26, .04	-.05, .21
$TILT$	Males	-.46, .36	-.44, .33	-.56, .27	-.39, .32	.04, .58
	Females	-.29, .14	-.37, .04	-.30, .07	-.35, .08	-.09, .37

28, 29, 30, respectively. In Table 28, the near-zero (and sometimes negative) r_c values obtained in most cases for the 25^{th} percentile speaker indicate that a consistent correlation across speakers could not be maintained for any IF feature. Nevertheless, as shown in Table 30, the 75^{th} percentile speakers did obtain moderate to high values (0.60 – 0.81) for the following IF features: OQa, $SQ_{20\text{-}80}$, OQ_{50}, OQ_{80}, and $TILT$ for male speakers, as well as NAQ for female speakers. This result suggests that for some speakers, the EGG may in fact contain information that is approximately equivalent to the aforementioned airflow features, even if the relation is speaker-specific. It is interesting to note that when comparing time-domain to DCT EGG features, there was no obvious trend as to which feature set led to more accurate estimates of the IF features. These feature sets appear to be similar in their overall ability to model glottal airflow features.

B.6 Conclusion

The results presented in the previous two sections demonstrate a complex, speaker-specific relationship between EGG and glottal airflow features. Possible explanations for the speaker-dependence of these relationships include imprecise placement of the EGG electrodes and variations in extra-laryngeal neck structures, both of which can affect the EGG signal [27]. While the results presented herein are specific to sustained utterances and to the EGG feature extraction methods employed, the lack of a consistent contact-airflow relationship across speakers puts into question the usefulness of EGG features as estimation targets for the glottal waveform characterization method proposed in this thesis. This is because changes in the EGG waveform that are not related to specific changes in the glottal airflow pattern are unlikely to have consistent, predictable acoustic effects, thus precluding the estimation of EGG features from the spectral envelope of the acoustic speech signal. Additionally, the limited size of available EGG corpora, described in Section B.1, is unsuitable for the goal of training a

large, speaker-independent statistical feature transformation model. Overall, unless a more intricate procedure for extracting features from the EGG signal that overcomes inter-speaker variation is found, the usefulness of the EGG signal within the context of the research goals of this thesis appears to be limited to providing voiced/unvoiced detection and GCI information, both of which can be well-approximated from a clean acoustic speech signal via well-established implementations of pitch and GCI detection algorithms [85, 109, 19, 102].

Table 28: Correlation coefficient r_c between IF and GMR airflow feature estimates. N_r denotes number of GMM components. GMR estimates obtained by transforming time-domain (tfeat) or DCT-based EGG features, respectively. 25^{th} percentile across speakers.

		$N_r = 2$		$N_r = 4$		$N_r = 8$		$N_r = 16$	
		tfeat	DCT	tfeat	DCT	tfeat	DCT	tfeat	DCT
OQ_{LF}	Males	-0.19	-0.07	0.02	0.04	-0.01	0.01	0.17	0.07
	Females	-0.12	-0.04	-0.10	-0.06	-0.05	0.06	0.00	0.10
α_m	Males	-0.25	-0.10	-0.19	-0.17	-0.06	0.01	0.05	0.04
	Females	-0.17	-0.06	-0.03	-0.06	-0.09	-0.03	-0.02	0.05
Qa	Males	-0.16	-0.09	-0.05	0.00	-0.07	0.05	0.02	0.17
	Females	-0.15	0.01	-0.06	0.12	-0.01	0.09	0.05	0.07
Rd	Males	-0.14	-0.18	-0.03	-0.15	-0.23	0.08	0.02	-0.06
	Females	-0.14	0.03	-0.05	0.05	0.04	0.00	0.08	0.12
NAQ	Males	-0.09	-0.27	-0.06	-0.06	-0.07	-0.04	-0.13	0.08
	Females	-0.15	0.21	-0.13	0.02	-0.07	0.12	0.20	0.27
ClQ	Males	-0.04	0.03	-0.07	-0.18	-0.11	-0.22	0.04	0.00
	Females	-0.09	-0.08	-0.09	-0.10	-0.12	-0.02	0.04	0.05
OQa	Males	0.00	-0.01	-0.21	0.00	0.05	0.10	0.15	0.21
	Females	-0.10	-0.17	-0.23	-0.16	-0.08	-0.06	0.09	0.04
SQ_1	Males	-0.24	-0.20	-0.09	-0.31	-0.20	-0.04	-0.05	0.08
	Females	-0.14	0.04	-0.05	-0.02	-0.07	-0.02	0.05	0.03
SQ_2	Males	-0.18	-0.25	-0.06	-0.16	-0.04	-0.02	-0.05	0.06
	Females	-0.10	-0.09	0.00	-0.14	0.01	-0.04	0.01	0.06
$SQ_{20\text{-}80}$	Males	-0.01	-0.02	0.12	0.01	-0.02	0.17	0.19	-0.03
	Females	-0.31	-0.06	-0.22	0.04	-0.06	0.05	0.06	0.21
OQ_{20}	Males	-0.11	-0.06	-0.05	-0.22	0.03	0.00	0.06	0.02
	Females	-0.28	-0.35	-0.26	-0.11	-0.10	-0.04	0.01	0.13
OQ_{50}	Males	-0.06	-0.30	-0.12	-0.20	0.11	0.08	0.26	0.08
	Females	-0.32	-0.21	-0.26	-0.15	-0.29	0.02	0.05	0.06
OQ_{80}	Males	-0.12	-0.13	0.13	0.08	0.08	-0.02	0.21	0.23
	Females	-0.32	-0.13	-0.16	-0.05	-0.07	-0.05	0.03	-0.04
$H1$-$H2$	Males	-0.14	-0.12	-0.17	-0.11	-0.08	-0.10	-0.09	-0.10
	Females	-0.11	-0.04	-0.10	-0.05	-0.07	-0.06	-0.04	-0.05
HRF	Males	0.02	-0.13	-0.02	0.10	-0.02	0.11	0.18	0.14
	Females	-0.08	-0.03	-0.07	0.07	-0.05	0.11	0.05	0.12
$TILT$	Males	0.01	-0.04	-0.12	0.11	-0.02	0.12	0.23	0.20
	Females	-0.10	-0.10	-0.08	-0.02	-0.03	0.04	0.05	0.17

Table 29: Correlation coefficient r_c between IF and GMR airflow feature estimates. N_r denotes number of GMM components. GMR estimates obtained by transforming time-domain (tfeat) or DCT-based EGG features, respectively. Median across speakers.

		$N_r = 2$		$N_r = 4$		$N_r = 8$		$N_r = 16$	
		tfeat	DCT	tfeat	DCT	tfeat	DCT	tfeat	DCT
OQ_{LF}	Males	0.17	0.32	0.26	0.31	0.19	0.29	0.33	0.34
	Females	0.00	0.15	-0.02	0.13	0.06	0.15	0.12	0.22
α_m	Males	-0.10	0.11	-0.03	0.00	0.11	0.14	0.27	0.21
	Females	-0.02	0.15	0.08	0.11	0.04	0.13	0.15	0.17
Qa	Males	0.12	0.02	0.03	0.15	0.13	0.17	0.20	0.24
	Females	-0.01	0.18	0.09	0.22	0.15	0.16	0.18	0.17
Rd	Males	0.08	0.06	0.15	0.04	0.02	0.19	0.15	0.18
	Females	0.02	0.17	0.08	0.22	0.12	0.16	0.15	0.19
NAQ	Males	0.10	-0.01	0.17	0.11	0.20	0.12	0.24	0.18
	Females	0.12	0.41	0.17	0.28	0.25	0.35	0.25	0.38
ClQ	Males	0.15	0.21	0.15	0.15	0.16	0.05	0.23	0.26
	Females	0.11	0.12	0.10	0.03	0.06	0.21	0.18	0.27
OQa	Males	0.30	0.16	0.15	0.16	0.22	0.30	0.44	0.37
	Females	0.08	0.04	-0.02	0.05	0.12	0.10	0.18	0.13
SQ_1	Males	0.09	0.09	0.22	0.23	0.30	0.20	0.38	0.37
	Females	0.02	0.25	0.08	0.09	0.11	0.16	0.23	0.27
SQ_2	Males	0.13	-0.01	0.13	0.11	0.05	0.24	0.19	0.24
	Females	0.00	0.22	0.12	0.11	0.09	0.19	0.13	0.24
$SQ_{20\text{-}80}$	Males	0.31	0.29	0.29	0.37	0.42	0.39	0.54	0.46
	Females	-0.02	0.16	0.11	0.22	0.04	0.23	0.22	0.42
OQ_{20}	Males	0.06	0.08	0.07	0.00	0.19	0.06	0.28	0.18
	Females	-0.04	-0.08	0.05	0.15	0.12	0.16	0.11	0.28
OQ_{50}	Males	0.34	0.09	0.25	0.24	0.26	0.22	0.41	0.38
	Females	-0.11	0.14	0.08	0.13	0.09	0.13	0.21	0.21
OQ_{80}	Males	0.17	0.22	0.29	0.38	0.33	0.37	0.44	0.48
	Females	0.01	0.09	0.18	0.25	0.21	0.27	0.26	0.40
$H1$-$H2$	Males	-0.03	-0.04	-0.04	-0.03	0.03	0.05	0.02	0.02
	Females	-0.04	0.00	-0.03	0.02	0.02	0.02	0.01	0.03
HRF	Males	0.18	0.21	0.27	0.21	0.25	0.30	0.40	0.34
	Females	0.07	0.15	0.07	0.19	0.06	0.23	0.18	0.24
$TILT$	Males	0.36	0.21	0.19	0.27	0.21	0.45	0.41	0.48
	Females	0.16	0.07	0.08	0.13	0.15	0.14	0.22	0.29

Table 30: Correlation coefficient r_c between IF and GMR airflow feature estimates. N_r denotes number of GMM components. GMR estimates obtained by transforming time-domain (tfeat) or DCT-based EGG features, respectively. 75^{th} percentile across speakers.

		$N_r = 2$		$N_r = 4$		$N_r = 8$		$N_r = 16$	
		tfeat	DCT	tfeat	DCT	tfeat	DCT	tfeat	DCT
OQ_{LF}	Males	0.52	0.56	0.56	0.46	0.57	0.57	0.58	0.56
	Females	0.12	0.27	0.20	0.25	0.31	0.31	0.35	0.33
α_m	Males	0.21	0.34	0.16	0.26	0.29	0.26	0.42	0.38
	Females	0.15	0.28	0.18	0.28	0.20	0.36	0.33	0.35
Qa	Males	0.41	0.19	0.38	0.34	0.25	0.35	0.44	0.46
	Females	0.09	0.30	0.18	0.39	0.29	0.33	0.39	0.39
Rd	Males	0.18	0.21	0.33	0.25	0.22	0.25	0.30	0.41
	Females	0.10	0.31	0.16	0.35	0.28	0.30	0.32	0.36
NAQ	Males	0.22	0.23	0.40	0.32	0.37	0.28	0.54	0.36
	Females	0.33	0.56	0.37	0.47	0.39	0.46	0.54	0.64
ClQ	Males	0.27	0.42	0.27	0.36	0.32	0.22	0.42	0.34
	Females	0.22	0.43	0.37	0.29	0.26	0.30	0.29	0.45
OQa	Males	0.62	0.47	0.37	0.46	0.40	0.48	0.57	0.64
	Females	0.22	0.12	0.12	0.23	0.38	0.21	0.50	0.42
SQ_1	Males	0.36	0.43	0.40	0.42	0.50	0.55	0.52	0.49
	Females	0.21	0.36	0.26	0.29	0.31	0.36	0.39	0.49
SQ_2	Males	0.31	0.39	0.33	0.37	0.28	0.40	0.46	0.53
	Females	0.21	0.33	0.20	0.25	0.25	0.26	0.30	0.43
$SQ_{20\text{-}80}$	Males	0.69	0.62	0.63	0.56	0.75	0.63	0.81	0.67
	Females	0.20	0.42	0.33	0.48	0.40	0.46	0.41	0.51
OQ_{20}	Males	0.26	0.19	0.35	0.21	0.32	0.37	0.47	0.32
	Females	0.18	0.08	0.27	0.48	0.26	0.36	0.36	0.47
OQ_{50}	Males	0.62	0.45	0.43	0.41	0.44	0.52	0.66	0.59
	Females	0.23	0.48	0.30	0.24	0.26	0.31	0.41	0.42
OQ_{80}	Males	0.48	0.66	0.59	0.60	0.61	0.55	0.71	0.72
	Females	0.25	0.27	0.44	0.53	0.43	0.56	0.49	0.65
$H1$-$H2$	Males	0.06	0.16	0.13	0.13	0.18	0.24	0.19	0.25
	Females	0.02	0.10	0.02	0.07	0.07	0.06	0.08	0.11
HRF	Males	0.50	0.58	0.53	0.49	0.52	0.50	0.57	0.59
	Females	0.23	0.29	0.29	0.34	0.26	0.36	0.39	0.42
$TILT$	Males	0.49	0.46	0.37	0.53	0.39	0.61	0.69	0.57
	Females	0.26	0.26	0.23	0.25	0.33	0.30	0.35	0.37

REFERENCES

[1] Airas, M., "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logop. Phoniatr. Voco.*, vol. 33, no. 1, p. 49, 2008.

[2] Airas, M., Pulakka, H., Bäckström, T., and Alku, P., "TKK Aparat ver. 0.3.1." [Online]. Available: `http://aparat.sourceforge.net`. [Accessed Mar. 15, 2010].

[3] Airas, M. and Alku, P., "Emotions in short vowel segments: Effects of the glottal flow as reflected by the normalized amplitude quotient," in *Affective Dialogue Systems* (André, E., Dybkjær, L., Minker, W., and Heisterkamp, P., eds.), vol. 3068 of *Lecture Notes in Computer Science*, pp. 13–24, Springer, 2004.

[4] Akande, O. O. and Murphy, P. J., "Estimation of the vocal tract transfer function with application to glottal wave analysis," *Speech Commun.*, vol. 46, no. 1, pp. 15–36, 2005.

[5] Alku, P., Airas, M., Backstrom, T., and Pulakka, H., "Group delay function as a means to assess quality of glottal inverse filtering," in *Proc. INTERSPEECH-2005*, (Lisbon, Portugal), pp. 1053–1056, 2005.

[6] Alku, P., Airas, M., and Story, B., "Evaluation of an inverse filtering technique using physical modeling of voice production," in *Proc. INTERSPEECH-2004*, (Jeju Island, Korea), pp. 497–500, 2004.

[7] Alku, P. and Vilkman, E., "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers.," *Folia Phoniatr. Logo.*, vol. 48, no. 5, pp. 240–54, 1996.

[8] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, no. 2-3, pp. 109–118, 1992.

[9] Alku, P., "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Am.*, vol. 112, no. 2, p. 701, 2002.

[10] Alku, P., Magi, C., Yrttiaho, S., Backstrom, T., and Story, B., "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3289–3305, 2009.

[11] Ananthapadmanabha, T. V. and Fant, G., "Calculation of true glottal flow and its components," *Speech Commun.*, vol. 1, no. 3-4, pp. 167–184, 1982.

[12] Arroabarren, I. and Carlosena, A., "Unified analysis of glottal source spectrum," in *Proc. INTERSPEECH-2003*, (Geneva, Switzerland), pp. 1761–1764, 2003.

[13] Backstrom, T., Airas, M., Lehto, L., and Alku, P., "Objective quality measures for glottal inverse filtering of speech pressure signals," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, vol. 1, pp. 897–900, 2005.

[14] Badin, P., Hertegård, S., and Karlsson, I., "Notes on the rothenberg mask," *STL-QPSR, KTH*, vol. 1, pp. 1–7, 1990.

[15] Baken, R. J. and Orlikoff, R. F., *Clinical measurement of speech and voice*. Singular, 2nd ed., 2000.

[16] Berg, J. W. V. D., "On the air response and the bernoulli effect of the human larynx," *J. Acoust. Soc. Am.*, vol. 29, pp. 626–631, 1957.

[17] Bozkurt, B., Doval, B., D'Alessandro, C., and Dutoit, T., "Zeros of z-transform representation with application to source-filter separation in speech," *IEEE Signal Process. Lett.*, vol. 12, no. 4, pp. 344–347, 2005.

[18] Brookes, D. M. and Chan, D. S. F., "Speaker characteristics from a glottal airflow model using robust inverse filtering," *Proc. Institute of Acoustics*, vol. 16, pp. 501–508, 1994.

[19] Brookes, M., "VOICEBOX: Speech processing toolbox for MATLAB." [Online]. Available: `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. [Accessed Mar. 15, 2010].

[20] Childers, D. and Krishnamurthy, A., "A critical review of electroglottography.," *Crit. Rev. Biomed. Eng.*, vol. 12, no. 2, pp. 131–61, 1985.

[21] Childers, D. and Lee, C., "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Am.*, vol. 90, p. 2394, 1991.

[22] Childers, D. G., "Glottal source modeling for voice conversion," *Speech Commun.*, vol. 16, no. 2, pp. 127–138, 1995.

[23] Childers, D. G., *Speech processing and synthesis toolboxes*. John Wiley and Sons, Inc., 2000.

[24] Childers, D. G., Hicks, D. M., Moore, G. P., and Alsaka, Y. A., "A model for vocal fold vibratory motion, contact area, and the electroglottogram," *J. Acoust. Soc. Am.*, vol. 80, no. 5, pp. 1309–1320, 1986.

[25] Childers, D. G. and Wong, C., "Measuring and modeling vocal source interaction," *IEEE Trans. Biomed. Eng.*, vol. 41, no. 7, pp. 663–671, 1994.

[26] Coleman, T. F. and Li, Y., "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on Optimization*, vol. 6, no. 2, pp. 418–445, 1996.

[27] Colton, R. H. and Conture, E. G., "Problems and pitfalls of electroglottography," *J. Voice*, vol. 4, no. 1, pp. 10–24, 1990.

[28] Cummings, K. and Clements, M., "Analysis of the glottal excitation of emotionally stressed speech," *J. Acoust. Soc. Am.*, vol. 98, pp. 88–98, 1995.

[29] Cummings, K. E. and Clements, M. A., "Glottal models for digital speech processing: A historical review and new results," *Digital Signal Processing: A Review Journal*, vol. 5, no. 1, pp. 21–42, 1995.

[30] Davis, S. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, pp. 357–366, Aug 1980.

[31] del Pozo, A. and Young, S., "The linear transformation of LF glottal waveforms for voice conversion," in *Proc. INTERSPEECH-2008*, (Brisbane, Australia), pp. 1457–1460, 2008.

[32] Deller, J., "Some notes on closed phase glottal inverse filtering," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 4, pp. 917–919, 1981.

[33] Devore, J. L., *Probability and Statistics for Engineering and the Sciences*. Pacific Grove, CA: Brooks/Cole, 4th ed., 1995.

[34] Dibazar, A. A., Narayanan, S., and Berger, T. W., "Feature analysis for automatic detection of pathological speech," in *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBS) Meeting*, vol. 1, pp. 182–183, 2002.

[35] Dinther, R. v., Veldhuis, R., and Kohlrausch, A., "Perceptual aspects of glottal-pulse parameter variations," *Speech Commun.*, vol. 46, no. 1, pp. 95–112, 2005.

[36] Doval, B., d'Alessandro, C., and Henrich, N., "The spectrum of glottal flow models," *Acta Acust. United Ac.*, vol. 92, no. 6, pp. 1026–1046, 2006.

[37] Dromey, C., Stathopoulos, E. T., and Sapienza, C. M., "Glottal airflow and electroglottographic measures of vocal function at multiple intensities," *J. Voice*, vol. 6, no. 1, pp. 44–54, 1992.

[38] El-Jaroudi, A. and Makhoul, J., "Discrete all-pole modeling," *IEEE Trans. Signal Process.*, vol. 39, pp. 411–423, feb 1991.

[39] Espy-Wilson, C. Y., Manocha, S., and Vishnubhotla, S., "A new set of features for text-independent speaker identification," in *Proc. INTERSPEECH-2006*, (Pittsburgh, PA, USA), pp. 1475–1478, 2006.

[40] Fant, G., "The LF-model revisited. transformations and frequency domain analysis," Q. Prog. Status Rep. STL-QPSR 2–3/95, Speech Transmission Laboratory, Royal Institute of Technology (KTH), 1995.

[41] Fant, G., Liljencrants, J., and Lin, Q., "A four-parameter model of glottal flow," Q. Prog. Status Rep. STL-QPSR 4/85, Speech Transmission Laboratory, Royal Institute of Technology (KTH), Stockholm, Sweden, 1985.

[42] Fernandez, R. and Picard, R. W., "Classical and novel discriminant features for affect recognition from speech," in *Proc. INTERSPEECH-2005*, (Lisbon, Portugal), pp. 473–476, 2005.

[43] Frohlich, M., Michaelis, D., and Strube, H. W., "SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals," *J. Acoust. Soc. Am.*, vol. 110, no. 1, pp. 479–88, 2001.

[44] Fu, Q. and Murphy, P., "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 2, pp. 492–501, 2006.

[45] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *NTIS order number PB91-505065*, 1990.

[46] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L., "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM documentation," *NTIS order number PB93-173938*, 1993.

[47] Gobl, C. and Chasaide, A. N., "Amplitude-based source parameters for measuring voice quality," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, (Geneva, Switzerland), pp. 151–156, 2003.

[48] Gobl, C. and Chasaide, A. N., "The role of voice quality in communicating emotion, mood, and attitude," *Speech Commun.*, vol. 40, pp. 189–212, 2003.

[49] Gold, B. and Morgan, N., *Speech and audio signal processing: processing and perception of speech and music*. Wiley, 2000.

[50] Granqvist, S., Hertegård, S., Larsson, H., and Sundberg, J., "Simultaneous analysis of vocal fold vibration and transglottal airflow: exploring a new experimental setup," *J. Voice*, vol. 17, no. 3, p. 319, 2003.

[51] Gudnason, J. and Brookes, M., "Voice source cepstrum coefficients for speaker identification," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, pp. 4821–4824, 2008.

[52] Guerin, B., Mrayati, M., and Carre, R., "A voice source taking account of coupling with the supraglottal activities," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, vol. 1, pp. 47–50, 1976.

[53] Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L., eds., *Feature Extraction: Foundations and Applications*. Springer Verlag, 2006.

[54] Hansen, J. and Patil, S., "Speech under stress: Analysis, modeling and recognition," in *Speaker Classification I* (Müller, C., ed.), vol. 4343 of *Lecture Notes in Computer Science*, pp. 108–137, Springer, 2007.

[55] Hanson, H. M., "Glottal characteristics of female speakers: Acoustic correlates," *J. Acoust. Soc. Am.*, vol. 101, no. 1, pp. 466–481, 1997.

[56] Hanson, H. M. and Chuang, E. S., "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data," *J. Acoust. Soc. Am.*, vol. 106, no. 2, pp. 1064–1077, 1999.

[57] Hardcastle, W. J. and Laver, J., *The handbook of phonetic sciences*. Wiley-Blackwell, 1999.

[58] Hayes, M. H., *Statistical digital signal processing and modeling*. John Wiley & Sons, 1996.

[59] Henrich, N., d'Alessandro, C., and Doval, B., "Spectral correlates of voice open quotient and glottal flow asymmetry: Theory, limits and experimental data," in *Proc. EUROSPEECH-2001*, (Aalborg, Denmark), pp. 47–50, 2001.

[60] Henrich, N., Doval, B., and Castellengo, M., "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. Acoust. Soc. Am.*, vol. 115, p. 1321, 2004.

[61] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[62] Hertegård, S., Gauffin, J., and Karlsson, I., "Physiological correlates of the inverse filtered flow waveform," *J. Voice*, vol. 6, no. 3, pp. 224–234, 1992.

[63] Hertegård, S., Larsson, H., and Wittenberg, T., "High-speed imaging: applications and development.," *Logop. Phoniatr. Voco.*, vol. 28, no. 3, pp. 133–139, 2003.

[64] Howard, D. M., "Variation of electrolaryngographically derived closed quotient for trained and untrained adult female singers," *J. Voice*, vol. 9, no. 2, pp. 163–172, 1995.

[65] Iseli, M. and Alwan, A., "An improved correction formula for the estimation of harmonic magnitudes and its application to open quotient estimation," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, vol. 1, pp. 669–672, 2004.

[66] Juang, B. H., Rabiner, L., and Wilpon, J., "On the use of bandpass liftering in speech recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 35, no. 7, pp. 947–954, 1987.

[67] Kain, A., *High Resolution Voice Transformation*. PhD thesis, Oregon Health and Science University, 2001.

[68] Kain, A. and Macon, M. W., "Spectral voice conversion for text-to-speech synthesis," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, vol. 1, pp. 285–288, 1998.

[69] Klatt, D. H. and Klatt, L. C., "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, pp. 820–857, 1990.

[70] Kounoudes, A., Naylor, P. A., and Brookes, M., "The DYPSA algorithm for estimation of glottal closure instants in voiced speech," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, vol. 1, pp. 349–352, 2002.

[71] Krishnamurthy, A. and Childers, D., "Two-channel speech analysis," *IEEE Trans. Acoust. Speech Signal Process*, vol. 34, no. 4, pp. 730–743, 1986.

[72] Laukkanen, A.-M., Vilkman, E., Alku, P., and Oksanen, H., "Physical variations related to stress and emotional state: a preliminary study," *J. Phonetics*, vol. 24, no. 3, pp. 313–335, 1996.

[73] Linville, S. E., "Glottal gap configurations in two age groups of women," *J. Speech Lang. Hear. R.*, vol. 35, no. 6, p. 1209, 1992.

[74] Lofqvist, A. and Mandersson, B., "Long-time average spectrum of speech and voice analysis.," *Folia Phoniatr.*, vol. 39, no. 5, pp. 221–229, 1987.

[75] Meyer, P., Schroeter, J., and Sondhi, M. M., "Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks," *IEEE Trans. Signal Process.*, vol. 39, no. 7, pp. 1493–1502, 1991.

[76] Milner, B. and Shao, X., "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 24–33, 2007.

[77] Ming, J., Hazen, T. J., Glass, J. R., and Reynolds, D. A., "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, 2007.

[78] Moore, E., Clements, M., Peifer, J., and Weisser, L., "Investigating the role of glottal features in classifying clinical depression," in *Proc., 25th Annual Conf. on Eng. in Medicine and Biology*, vol. 3, pp. 2849–2852, 2003.

[79] Moore, E., Clements, M. A., Peifer, J. W., and Weisser, L., "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 1, pp. 96–107, 2008.

[80] Moore, E. and Torres, J., "Improving glottal waveform estimation through rank-based glottal quality," in *Proc. INTERSPEECH-2006*, (Pittsburgh, PA, USA), pp. 1694–1697, 2006.

[81] Moore, E. and Torres, J., "A performance assessment of objective measures for evaluating the quality of glottal waveform estimates," *Speech Commun.*, vol. 50, no. 1, pp. 56–66, 2008.

[82] Mouchtaris, A., Van der Spiegel, J., and Mueller, P., "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 952–963, 2006.

[83] Murty, K. S. R. and Yegnanarayana, B., "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Proc. Let.*, vol. 13, no. 1, pp. 52–55, 2006.

[84] Nadeu, C., Hernando, J., and Gorricho, M., "On the decorrelation of filter-bank energies in speech recognition," in *Proc. EUROSPEECH '95*, (Madrid, Spain), pp. 923–926, 1995.

[85] Naylor, P. A., Kounoudes, A., Gudnason, J., and Brookes, M., "Estimation of glottal closure instants in voiced speech using the dypsa algorithm," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 1, pp. 34–43, 2007.

[86] Neiberg, D., Elenius, K., and Laskowski, K., "Emotion recognition in spontaneous speech using gmms," in *Proc. INTERSPEECH-2006*, (Pittsburgh, PA, USA), pp. 809–812, 2006.

[87] Ní Chasaide, A. and Gobl, C., *Voice source variation*, ch. 14, pp. 427–461. The Handbook of Phonetic Sciences, Blackwell Publishers, 1997.

[88] Ozdas, A., Shiavi, R. G., Silverman, S. E., Silverman, M. K., and Wilkes, D. M., "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 9, pp. 1530–1540, 2004.

[89] Ozdas, A., Wilkes, D. M., Shiavi, R. G., Silverman, S. E., and Silverman, M. K., "Analysis of fundamental frequency for near term suicidal risk assessment," in *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, vol. 3, pp. 1853–1858, 2000.

[90] Paliwal, K. K., "Decorrelated and liftered filter-bank energies for robust speech recognition," in *Proc. EUROSPEECH '99*, (Budapest, Hungary), pp. 85–88, 1999.

[91] Plumpe, M. D., Quatieri, T. F., and Reynolds, D. A., "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 5, pp. 569–585, 1999.

[92] Pulakka, H., "Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography," Master's thesis, Helsinki University of Technology, 2005.

[93] Quatieri, T. F., *Discrete-Time Speech Signal Processing: Principles and Practice*. Upper Saddle River, NJ: Prentice Hall PTR, 2002.

[94] Rabiner, L. R. and Juang, B. H., *Fundamentals of speech recognition*. Prentice hall, 1993.

[95] Rabiner, L. R. and Schafer, R. W., *Digital Processing of Speech Signals*. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.

[96] Reynolds, D. A., "Speaker identification and verification using gaussian mixture speaker models," *Speech Commun.*, vol. 17, no. 1-2, pp. 91–108, 1995.

[97] Reynolds, D. A. and Rose, R. C., "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, 1995.

[98] Rothenberg, M., "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Am.*, vol. 53, p. 1632, 1973.

[99] Rothenberg, M. and Mahshie, J., "Monitoring vocal fold abduction through vocal fold contact area," *J. Speech Lang. Hear. R.*, vol. 31, no. 3, pp. 338–351, 1988.

[100] Saito, S., Fukuda, H., Kitahara, S., and Kokawa, N., "Stroboscopic observation of vocal fold vibration with fiberoptics.," *Folia Phoniatr.*, vol. 30, no. 4, p. 241, 1978.

[101] Sapienza, C., Stathopoulos, E., and Dromey, C., "Approximations of open quotient and speed quotient from glottal airflow and egg waveforms : Effects of measurement criteria and sound pressure level," *J. Voice*, vol. 12, no. 1, p. 31, 1998.

[102] Sjölander, K., "The snack sound toolkit ver. 2.2.10." [Online]. Available: `http://www.speech.kth.se/snack/index.html`. [Accessed Mar. 15, 2010].

[103] Sodersten, M. and Lindestad, P. A., "Glottal closure and perceived breathiness during phonation in normally speaking subjects," *J. Speech Lang. Hear. R.*, vol. 33, no. 3, p. 601, 1990.

[104] Sturmel, N., D'Alessandro, C., and Doval, B., "A comparative evaluation of the zeros of z-transform representation for voice source estimation," in *Proc. INTERSPEECH-2007*, (Antwerp, Belgium), pp. 558–561, 2007.

[105] Sun, R., Moore, E., and Torres, J. F., "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, pp. 4509–4512, april 2009.

[106] Sung, H. G., *Gaussian Mixture Regression and Classification*. PhD thesis, Rice University, 2004.

[107] Svec, J. G. and Schutte, H. K., "Videokymography: high-speed line scanning of vocal fold vibration.," *J. Voice*, vol. 10, no. 2, pp. 201–205, 1996.

[108] Swerts, M. and Veldhuis, R., "The effect of speech melody on voice quality," *Speech Commun.*, vol. 33, no. 4, pp. 297–303, 2001.

[109] Talkin, D., *A robust algorithm for pitch tracking (RAPT)*, ch. 14, pp. 495–518. Speech coding and synthesis, Amsterdam, NL: Elsevier Science, 1995.

[110] Teager, H., "Some observations on oral air flow during phonation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 5, pp. 599–601, 1980.

[111] Teager, H., "Evidence for nonlinear sound production mechanisms in the vocal tract," in *NATO ASI: Speech production and speech modeling*, pp. 1–21, 1989.

[112] Theodoridis, S. and Koutroumbas, K., *Pattern Recognition*. San Diego, CA: Elsevier, 1999.

[113] Thomas, M. R. P. and Naylor, P. A., "The sigma algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, pp. 1557–1566, nov. 2009.

[114] Torres, J. and Moore, E., "Performance evaluation of glottal quality measures from the perspective of vocal tract filter consistency," in *Proc. INTERSPEECH-2007*, (Antwerp, Belgium), pp. 66–69, 2007.

[115] Torres, J. F., Moore, E., and Bryant, E., "A study of glottal waveform features for deceptive speech classification," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, pp. 4489–4492, 2008.

[116] Varadarajan, V. S. and Hansen, J. H. L., "Analysis of lombard effect under different types and levels of noise with application to in-set speaker id systems," in *Proc. INTERSPEECH-2006*, (Pittsburgh, PA, USA), pp. 937–940, 2006.

[117] Veeneman, D. and BeMent, S., "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Trans. Acoust. Speech Signal Process*, vol. 33, pp. 369–377, apr 1985.

[118] Veldhuis, R., "The spectral relevance of glottal-pulse parameters," in *Proc. IEEE Int. Conf. Acous. Spch. Sig. Process.*, vol. 2, pp. 873–876, 1998.

[119] Waaramaa, T., Laukkanen, A., Alku, P., and Väyrynen, E., "Monopitched expression of emotions in different vowels.," *Folia Phoniatr. Logo.*, vol. 60, no. 5, pp. 249–255, 2008.

[120] Walker, J. and Murphy, P., "A review of glottal waveform analysis," in *Progress in Nonlinear Speech Processing*, vol. 4391 of *Lecture Notes in Computer Science*, pp. 1–21, Springer, 2007.

[121] Wong, D., Markel, J., and Gray, A., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 4, pp. 350–355, 1979.

[122] Yan, Y., Ahmad, K., Kunduk, M., and Bless, D., "Analysis of vocal-fold vibrations from high-speed laryngeal images using a hilbert transform-based methodology," *J. Voice*, vol. 19, no. 2, pp. 161–175, 2005.

[123] Yan, Y., Damrose, E., and Bless, D., "Functional analysis of voice using simultaneous high-speed imaging and acoustic recordings," *J. Voice*, vol. 21, no. 5, pp. 604–616, 2007.

[124] Yoon, T., Zhuang, X., Cole, J., and Hasegawa-Johnson, M., "Voice quality dependent speech recognition," in *International Symposium on Linguistic Patterns in Spontaneous Speech*, (Taipei, Taiwan), pp. 77–100, Academia Sinica, 2006.

[125] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.

[126] Young, S., "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Processing Magazine*, vol. 13, p. 45, sep 1996.

[127] Zheng, N., Lee, T., and Ching, P. C., "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Proc. Let.*, vol. 14, pp. 181–184, march 2007.