STRATEGIES AND TOOLS FOR THE SONIFICATION OF PROSODIC DATA: A COMPOSER'S PERSPECTIVE

Fabio Cifariello Ciardi

Conservatorio "F.A. Bonporti" di Trento e Riva del Garda
Via S. Giovanni Bosco, 4
Trento, Italy
fabio.cifariellociardi@conservatorio.tn.it

ABSTRACT

Does it make sense to sonify information that refers to an already audible phenomenon such as prosodic data? In order to be useful, a sonification of prosody should contribute to the comprehension of paralinguistic features that may not otherwise attract the attention of the listener. Within this context, this paper illustrates a modular and flexible framework for the reduction and processing of prosodic data to be used for enhancing the perception of a speaker's intention, attitude and emotions. The model uses speech audio as input and provides MIDI and MusicXML data as output so allowing samplers and notation software to auralize and display the information. The described architecture has been subjectively tested by the author over many years in compositions for solo instruments, ensembles and orchestra. Two outcomes of the research are discussed: the advantages of an adaptive strategy for data reduction, and the auditory display of the deep pitch and temporal structures underlying prosodic processing.

1. INTRODUCTION

Various sonification paradigms have been implemented to convey information about dynamic systems in which vibrations are too low to be heard directly. More rarely, sonification techniques have been used to enhance the meaning of data collected from phenomena that are already audible to humans. The idea of representing sound through sound makes sense only if an augmented perception of the original source is proved to be useful for a better understanding and interpretation of the data [1].

A remarkable example within this framework might be represented by the prosodic flow of speech (i.e. duration, accent, volume, tempo, pitch and contour variations in the vocal signal). Several languages from Africa and East Asia make use of intonation to transmit meaning, while others use stress to differentiate ambiguous words. Yet, all adopt prosodic features to express thoughts, feelings and behavioral information well beyond words. So, do we really need to sonify prosody? Prosody typically conveys emotions or attitudes that can confirm or alter the interpretation of the meanings of an utterance. However, a one-to-one mapping between given prosodic features of speech and a particular pragmatic interpretation is not always reliable. Which is to say that lexical and prosodic meanings can be perceived as incongruent.

From a linguistic standpoint, the auditory display of elements of prosody that are not self-evident can be useful when discourse context may distort the listener's expectations for unique speaker meanings [2]. This is the case when external factors such as social norms or listener expectations pull vocal expressions in a certain direction, while physiological processes push the same vocal expressions in a different direction. Likewise, if prosodic cues are used across cultures to convey emotions [3], the ability to control and decode emotional prosody also depends on culture-specific cues [4]. Finally, today's spoken language interaction technologies rely on feature classification carried out using machine learning techniques. However, acoustic dimensions of prosody such as pitch, energy, timing and intensity are still needed when trying to isolate emotion specific information in the speech signal [5, 6].

From a musical perspective, composers may sonify prosody by addressing different perceptive components of the phonatory phenomenon. Surface prosodic structure may be made clearer by making the speech-melody and its source concurrently audible. When this is the case, pitch and rhythm are processed separately. On the one hand f0 tracking algorithms are used to highlight the profile of the fundamental frequency of the vowels and on the other the boundaries of the rhythmic units of speech are detected by comparing transient and steady parts of the auditory signal. The same procedure may be used to sonify speech intentions and to make audible the inner expressiveness of the speaker when the original (audio and/or video) source cannot be synchronously monitored. In this case longer parts of a spoken discourse need to be considered. In order for more deep-seated roots of prosody to be found the spectral data must be first resampled or mapped into a low-dimensional subspace and then clustered according to a given criterion (e.g. similarity indexing).

A tangible interest in the transcription of prosody for acoustic instruments is documented by letters, sketches and scores written by the 18th-century inventor Joshua Steele [7], and the 19th-century composer Leoš Janáček [8]. In the last half century, works in this field have been carried out by both researchers and composers. The most prominent examples include the 'musical naturalism' of François-Bernard Mâche's *Le son d'une voix* (1964); the transcriptions of the f0 contour carried out by Scott Johnson, Steve Reich and Steve Vai, and the 'synthrumentation' technique developed by Clarence Barlow in the 1980s; the Peter Ablinger's 'phonographs' in the 1990s; and more recently, the computer-aided orchestration tools used by Jonathan Harvey and several other composers (see [9, 10] for reviews). Among them, the composer PerMagnus Lindborg uses a modular approach similar to that proposed here to integrate the musicality of historical voices such as Mao Zedong and Olof Palme into chamber music scores [11]. Compared to Lindborg's computer assisted analysis and composition techniques, tools and strategies discussed in this paper are not limited to the transcription of the vocal line as a melody, but

https://doi.org/10.21785/icad2021.041

aim to provide a model for prosodic data sonification. That is to say a flexible framework for the reduction and processing of prosodic data to be used for enhancing its perception and facilitating its understanding.

2. SONIFYING PROSODY

The suggested approach uses speech audio as input and involves a three steps analysis-processing-synthesis procedure in the order in which it appears in the conventional signal chain.

2.1. Analysis

At this stage low-level spectral features are extracted at a frame-level. A fixed frame rate is deployed regardless of whether the signal is voiced or unvoiced (i.e., whether the audio in the frame has a harmonic structure or not, as in vowels). The data is organized into a tree structure where each leaf represents a coherent group of partials, as illustrated in Fig. 1.

```
((4700 840 60 60) (5300 840 60 40) (5600 840 60 40) (5900 840 60 80) (6200 840 60 40) (6600 840 60 60) (7100 840 60 40) (9300 840 60 40))
((4600 900 180 60) (5300 900 180 40) (5800 900 180 60) (6200 900 180 40) (6500 900 180 80) (7000 900 180 60) (9000 900 180 40))
((4900 1080 120 60) (6100 1080 120 60) (6700 1080 120 20) (7300 1080 120 20) (8000 1080 120 40))
((4800 1200 60 60) (6000 1200 60 80) (6700 1200 60 60) (7200 1200 60 80) (7600 1200 60 40) (8000 1200 60 40))
((4800 1260 60 60) (6000 1260 60 80) (6700 1260 60 80) (7200 1260 60 80) (7600 1260 60 40) (7900 1260 60 40) (8200 1260 60 40))
((4700 1320 120 60) (5800 1320 120 60) (6300 1320 120 20) (6700 1320 120 40) (7200 1320 120 60) (7600 1320 120 20) (7900 1320 120 20))
((4500 1440 60 60) (5700 1440 60 80) (6400 1440 60 40) (7000 1440 60 60))
((4400 1500 60 40) (5700 1500 60 60) (6000 1500 60 20) (6300 1500 60 20) (7000 1500 60 20))
((5700 1560 60 40))
((5300 1620 60 40) (5700 1620 60 40) (6000 1620 60 20) (6500 1620 60 40) (7200 1620 60 40) (8600 1620 60 20))
```

Figure 1: Example of the parametric representation of the data.

Each leaf node describes a single partial within fixed temporal limits with a given amplitude and frequency. The information about each partial includes its frequency in midicents (a midicent is a unit representing one cent of the usual MIDI pitch unit, that is, a half-tone), onset and duration in milliseconds, and amplitude coded as midi velocity from zero to 127. The higher is the resolution of the analysis the larger is the tree. This parametric representation is common to all modules of the system. It can be easily converted into a stream of MIDI data and sent to a synthesizer or sampler so allowing a smooth user interaction at each step of the sonification process.

The phase has not been considered in the parametric representation of the data for two reasons. Firstly, the aim of the proposed framework is to sonify prosodic information by transforming it into music performed by real musicians who are not able to make explicit phase control. Secondly, although the relative phase relationship within a critical band has been considered perceptually important [12], there is no evidence that the phase does in reality play a role in prosody. In the following examples[1] the first 12 partials of a male-voice excerpt [audio example n.1a] (Barack Obama saying, "How did this become such a partisan issue?") has been detected by partial tracking analysis and then synthesized by using inverse FFT [audio and SDIF example n.1b]. As expected, phase unwrapping [audio and SDIF example n.1c] is negligible in terms of timbre as compared to the original.

[1] Examples linked to this document are available at https://www.dropbox.com/sh/0zpkwpvn74aq12s/AADGU95yH0yF6t1QNDBmXh8Qa?dl=0. Audio clips are in single-channel WAV format sampled 44.1 kHz and normalized at −3 dB relative to 16-bit amplitude resolution. SDIF files use the 1TRC frame type [35].

2.2. Processing

At the processing stage, tradeoffs between resolution and quantization are necessary to assess the intrinsic dimensionality of data.

A raw speech signal is complex with its properties changing continuously. Its variable spectral content is the result of the complex way our physiological structures interact with the airflow supplied by the lungs. Laryngeal modulations, stationary or mobile vocal tract configurations contribute to the morphology of speech phonemes, giving shape to the words we hear. When compared to speech, prosodic information has a much lower dimensionality, but its variability is tightly bound to the number of syllables that speakers are able to produce. Experimental evidence shows that the syllabic rate across languages varies between 5 and 8 syllables per second (syll./s) in subjects asked to produce neither fast nor careful speech [13]. On the other hand, when subjects were encouraged to read a text in their native language ‘faster than normal', the results show mean values between about 8 syll./s for English and German and 11 syll./s for French speakers [14]. In musical terms, this means that to transcribe speech using just one sound per syllable, mean durations will vary between 200 and 91 ms (i.e., between a sixteenth note and a thirty-second note played at about 82 beats per minute). For instrumental music these constraints are manageable: speech segments played at such a speed can be performed by many acoustic instruments. The problem is that one sound per syllable is only sufficient to stress the surface commonalities between the prosody sonification of a spoken utterance and its source.

Apart from practical considerations, several issues underscore the need for data reduction in the auditory display of prosodic features. From a phonetical point of view, the characteristics of the signals can change significantly, but not all analysis frames need to be equally short. In automatic speech recognition a fixed frame rate with a window length of 20–30 ms is insufficient when classification and annotation tasks need to be solved [15]. On the other hand, in a prosodic data sonification system the detection of phonemic onsets and transitions aims at a different goal: assessing syllables boundaries so as to be able to extract prosodic rhythm from the speech signal. Example 2 [audio and SDIF] presents the inverse FFT reconstruction of the first 12 partials of audio example n.1a using 23 ms hop size. Despite the slight deterioration of the audio signal, the prosodic features are preserved.

The need for information reduction makes sense from a perceptual perspective as well. We constantly try to reduce the cognitive load by applying some sort of principal component analysis. For example, the perception of a changing pitch requires some minimal amount of frequency change as a function of time, otherwise a static pitch is perceived [16]. Finally, a strategy for data clustering is mandatory in the extraction of constants or slowly moving variables, which are often needed during the compositional integration of prosodic sonification.

2.3. Synthesis

In the third step of the procedure the prosodic data must be mapped onto an acoustic signal. Here, different sonification options must be made available to achieve the fine-tuning of the sonification system.

The first and more trivial solution is to synthetize each node of the data tree into sound without further reductions. A

second option includes the conversion of the tree into a MIDI stream to be played through a sampler. Example 3 [audio and MIDI] presents the MIDI data of SDIF example n.2 played with marimba samples. A logarithm curve has been used to map linear amplitudes into MIDI velocities. This option is even more important than the previous one because it lets the composer evaluate the optimal tradeoff between the informative add-on of a chosen timbre and its suitability for the compositional goals of the project. The cognitive processing elicited by the sound of a well-known musical instrument should not be underestimated. As stated by Craik [17] information is related conceptually to relevant pre-existing schematic knowledge, meaning that familiarity with sound source categories fosters a deep processing of the sonic environment experienced by the listener. Familiar instrumental sounds activate semantic, visual and even sensorimotor networks and these influence the way we attract attention and intention toward sounds [18].

The last option available in the synthesis step implies the conversion of the processed data into musical notation to be played by conventional instruments. At this stage, the dataset is further reduced to fulfill notational and instrumental constraints.

In natural speech, variations in tempo and duration depend on morphological, articulatory, language-dependent, emotional and contextual constraints. Rhythmic irregularities are the norm, rather than the exception: the longer the utterances, the more difficult is to balance the accuracy and playability of its prosodic sonification. On one hand, scoring these variations using fast-moving tempos allows the 'taming' of notation complexity. Yet, this solution is only truly feasible only for solo performances. When the execution in sync with the speech source is required or when more players are involved, rapid tempo changes are difficult to perform, even under the baton of a conductor. On the other hand, the more the tempo is kept constant, ever more complex irregular tuplets are needed to render the fluency of the prosody. In this respect, a lack of expertise on the issue can easily make the sonification impossible to be performed by humans.

3. PROSODY SONIFICATION WITH OPEN MUSIC

The presented modular approach has been implemented using Open Music, an object-oriented programming environment for music composition, analysis and research developed at IRCAM in the late nineties [19]. This Lisp-based language allows composers to develop functional processes generating or transforming musical data and to execute them locally by demand-driven evaluations.

The proposed sonification system uses the *partial-tracking* and *chord-seq-analysis* functions available in Open Music om-pm2 library [20]. The modules track sinusoidal frames in an audio file and returns a SDIF sound description file which can be represented as a data tree. All processing modules have been developed by the author. They implement algorithms to navigate, filter, quantize and analyze the data structure. In the final step, conversion functions make it possible to export data in MIDI and MusicXML facilitating samplers and notation softwares to auralize and display the information. The procedure has been tested using subjective evaluation experiments on realistic scenarios including challenging data input and validation exercise over many years. Several works scored for solo instruments [21], ensembles of various sizes [22], and orchestra [23] have been composed by the author using this sonification architecture.

Most of the time, the original speech signal and its prosodic sonification is diffused in parallel by providing ad-hoc cueing to the performer or the conductor. They could also run serially one after the other as in *Cupio Dissolvi* for pre-recorded voice and eight contrabasses [24]. The sonification is often coupled with visual information correlated to the phonetic material: the orator's face and gestures –with or without the corresponding audio content – or a combinatorial interplay between textual, pictorial and auditory cues (e.g. *Background checks* for pre-recorded voice, video and orchestra [23]).

There are two main outcomes of this research: an adaptive strategy for data reduction, and the emergence of the deep pitch and temporal structures underlying prosodic processing.

3.1. An adaptive strategy for data reduction

Speech signals are non-stationary but exhibit quasi-stationary behavior in shorter durations. However, certain acoustic attributes of the speech signal can be manifested in very short time intervals. A 1856ms speech segment analyzed using a fixed 23ms frame with a maximum of 12 partials is represented by 491 nodes of information. If the output of the sonification is intended to provide material for the composition of instrumental music, dimensionality reduction must be applied. In the time domain, this can be achieved using a variable or fixed resolution approach.

In the first case, analysis frames are downsampled dynamically: more frames are used for rapidly changing segments, and fewer for steady-state segments. However, this approach usually fails if only acoustic features are considered since noises and recording conditions can give rise to unpredictable changes in spectral energy. To overcome this problem machine learning technologies and probabilistic methods are widely implemented. An example of this approach is the Munich AUtomatic Segmentation web service (MAUS) [25] developed at the Institute of Phonetics and Speech Processing of the University of Munich. MAUS is based on probabilistic models: it calculates a phonetic segmentation and labeling for several languages based on the speech signal and the phonological transcript. The transcript is transformed into a probabilistic pronunciation graph which is then time-aligned to the speech signal. The application automatically subdivides the speech signals into segments according to the boundaries of each phoneme, thus reducing the number of frames to be analyzed. Apart from computational cost, similar segmentation approaches are not considered in this context for two reasons. Firstly, they are language dependent. Secondly, we must bear in mind that the analysis algorithm we use computes one sinusoidal chord for each 'slice' of the partitioned signal. Therefore, by using a phonetic segmentation as input, each segment will be represented by a single cluster of partials only. This tradeoff is acceptable if the goal is to sonify prosody on a phoneme-by-phoneme basis, but it is not enough when prosodic variation within a single phoneme (e.g. tonal glides in vowels) has to be considered.

Another way to downsample the time-domain component of the data is to use a fixed resolution approach. In this case, resolution is initially kept sufficiently high to detect any perceptible prosodic detail. Subsequently, the size of the data tree is progressively reduced by means of different kinds of filters and clustering operations. Here, the notions of

stylization may prove particularly valuable in shrinking prosodic features. The phonetician Pit Mertens [26] has referred to stylization as a procedure that modifies the pitch contour of an utterance into a simplified form, with the aim of preserving that part of melodic information which has a meaning in speech communication. The concept can be usefully extended to spectral contours by (1) calculating the centroid and the harmonic content of each analysis frame; (2) evaluating the variation in their moving average; and (3) filtering out frames below user-defined and context-dependent thresholds.

A similar approach has been employed in decreasing the duration resolution represented in each node to reduce the complexity of the musical notation output. In this case, durational values are firstly quantized using small binary subdivisions (e.g. thirty-second notes) [audio and video example 4a]. Secondly, the binary subdivisions are further quantized using different tuplets depending on compositional constrains (Fig.2) [video example 4b].

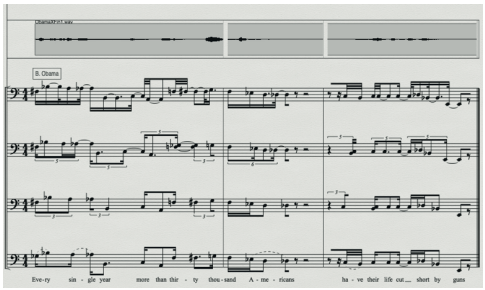

Figure 2: Progressive rhythm quantization of an intonation profile.

In the frequency domain, pitch resolution needs to be adjusted when the sonification is meant to be performed on acoustic instruments that do not allow continuous frequency changes (e.g. piano). The same constraints have also to be applied to instruments that permit some sort of microtonality (e.g. woodwinds), but only after long and time-consuming practice. In the author's experience, the lack of continuity between consecutive nodes of prosodic information is not as detrimental as it might seem at first glance. If the sonic articulation is fast enough, the Gestalt principles of auditory grouping can induce continuity illusions based on the limited spectral resolution of the auditory system [27]. Example 5 presents the inverse FFT reconstruction of the first 12 partials of audio example n.1a analyzed using 23 ms hop size with a twelve-tone quantization of the pitch and without [audio and SDIF example n.5b, audio examples n.5c and n.5d using marimba samples only].

Finally, amplitude resolution and quantization have to be reduced if sonifying prosody for live musicians. In the author's experience, several phonemes can be recognized even when the amplitude of speech partials is heavily quantized. In audio example n.6a the transcription of example n.1 is realized by appling a sixteenth-step amplitude quantization, whereas a three-step quantization only is used in example n.6b and n.6c. In performance practice, however, a trade-off between the variability of the data and instrumental constraints is difficult to achieve because of the subjective interpretation of the dynamic signs provided by traditional music notation. This is particularly true when (1) language or cultural conventions narrow the average dynamic range of speech; and (2) when prosodic features are correlated with spectral changes in the timbre of the source voice (e.g. vowel quality).

Summing up the above procedures, the modular design of the model provides an adaptive strategy to balance accuracy and playability in different instrumental settings: from the precise but complex notation for solo players (Fig. 3) [audio example n.7a, 7b], to the lower-resolution lattice which is necessary when scoring for large ensembles or orchestra (Fig. 4) [audio and video example n.7c].

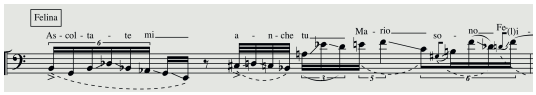

Figure 3: Example of prosodic transcription for solo trombone form *Appunti per Amanti Simulanei* for trombone, intonarumori and electronics [28].

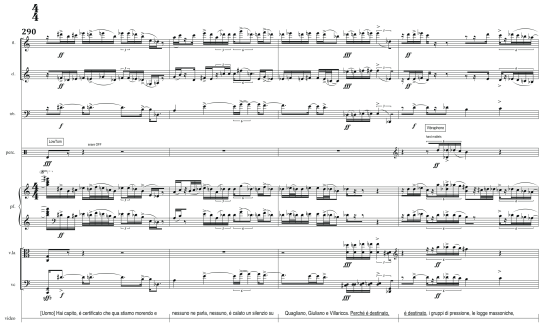

Figure 4: Example of prosodic transcription for ensemble from *Voci Vicine* Passion in 4 parts for journalist, video, ensemble and electronics [36].

The strategy is adaptive because it enables the composer to use different levels of resolution in the sonification of the prosodic data, depending on contextual and compositional conditions. Surface prosodic structures can be embedded into a single node of information per syllable using a drastic data reduction. When a finer resolution is needed more nodes may be used without changing the architecture of the system information.

3.2. Deep pitch and temporal structures in prosody

The main goal of data reduction is not just about making the transcription of prosodic features possible. Downsampling and quantization procedures are valuable only if the results convey information that is harder to interpret in normal speech-listening conditions. When this is the case, the reduction procedures do not compromise the meaningfulness of the source. On the contrary, it rather helps to enhance the perception of the prosodic prominences that the speaker uses to articulate the syntactic and emotional flow of the discourse.

To be effective prosodic punctuation has to be reduced when compared to the surface variation in rhythm, pitch contour and amplitude that we use to articulate an utterance. This has to do with the perceptive efficiency of the human auditory system: we can attribute a perceptual salience to items that are more prominent or emotionally striking only by ignoring those that are considered unremarkable.

Prosodic prominence emerges from the concurrent and coherent change of several acoustic dimensions. In this respect its meaning diverges from the most trivial concept of

musical accent. In music theory accentuation is mostly linked to amplitude variations, while a prominence is the sum of different factors such as loudness, length and pitch.

In the linguistic expressions of many languages, prominences (i.e. stress) are used to emphasize a given phoneme or syllable in speech. In some cases, they have a syntactic function in the interpretation of ambiguous words. In Italian, for example, a stress on the last or first syllable is what makes 'ce l'ho' (I have it) different from 'cielo' (sky). Tonal languages (e.g. Mandarin Chinese) convey different lexical meanings by the variation of fundamental frequency within the time span of a single syllable. In a non-tonal language like English prominences are used to clarify or disambiguate the semantics of an utterance. The primary cue of what is termed stress is pitch prominence [29]. In pitch-accent languages such as Japanese one single contour type is used for the realization of the accent. Yet in most non-tonal languages, they serve to make some parts of an utterance acoustically and perceptually more prominent and highlight parts of the information against the unaccented material.

Particular intonation contours may reveal the conscious or unconscious feelings of the speaker about what they are saying, especially with respect to affect (e.g. the emotion about some referent or proposition) [30]. In this context, the sonification of pitch accents through the mediation of music notation has some advantages if compared with the standard for labeling the prosody of speech, ToBI (i.e. tones and break indices) [31]. The ToBI system analyzes prominences using only two pitch targets, H (high) and L (low) associated with emphasis and with the ends of prosodic units. Instead, the sonification of prosody provides the listener with an augmented sonic representation of the entire hierarchical pitch structure that supports speech. Prominences that mark syntactic boundaries are positioned at the bottom of the hierarchy and might not be emphasized by any compositional device. Those that help the listener to focus on new information may be placed in the middle of the hierarchical structure and possibly slightly stressed while a pitch accent that slightly highlights one word among others – without seizing the attention of many listeners – might be situated at the top of the structure.

By highlighting this last class of prominences the sonification might contribute to the comprehension of the message communicated by the speaker even beyond their intention. For example, Fig 5 [video example n.8] shows an excerpt from the speech-melody of Barack Obama addressing the Arab students of the University of Cairo in 2009. Each segment of the utterance excerpt is emphasized by a pitch accent, but only one ("they've taught at our universities") is stressed by a pitch above A3. Considering the specificity of his live audience, it is worth considering that soaring pitch accent as not being casual.

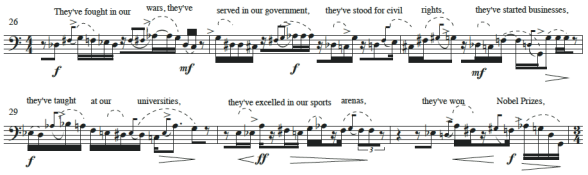

Figure 5: Excerpt from *Piccoli Studi sul Potere - Obama 06_04_2009* for cello and video [21].

In sonifying prosodic data, the composer may operate beyond the simple transcription of the speech-melody by exploring at least two directions. The first is to probe the pitch structures of prosody by emphasizing the perception of particular intonation contours. The second is to enhance multiple levels of stress through appropriate orchestration solutions. This is shown in the author's *Background Checks* [23] where the sonification of the prosodic features embedded in another speech by Obama [32] has been scored for orchestra. In the excerpt shown in Fig. 6 [audio example n.9] different layers of prominences have been sonified by strings chords of different perceptual 'weights' with the aim of rendering the seductive rhythm of his voice.

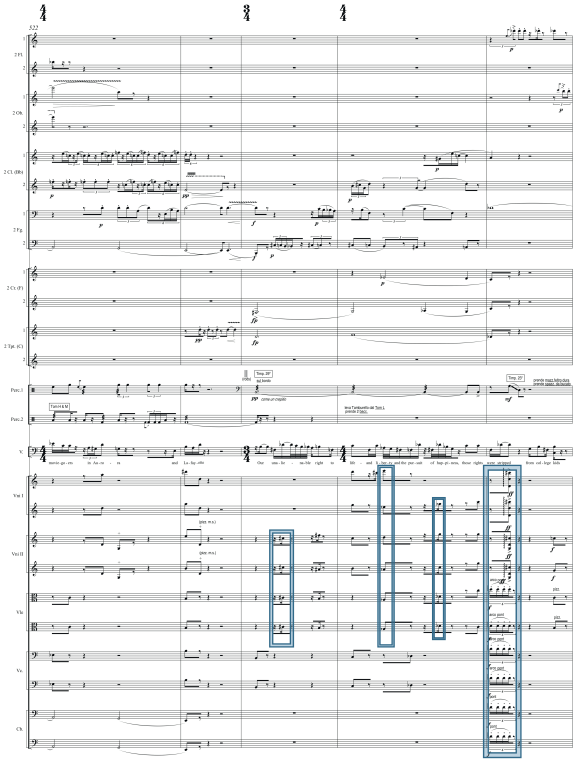

Figure 6: Orchestral sonification of different layers of prominences in the utterance "our unalienable right to life, and liberty, and the pursuit of happiness, those rights were stripped" from *Background Checks* for pre-recorded voice, video and orchestra [23].

4. DISCUSSION

There has been discussion about whether the goals of scientific and musical sonifications (e.g. data-driven compositions) are mutually exclusive or potentially inclusive [33]. In this respect, a prerequisite condition for communicating prosodic data within a musical context is a non-trivial dimensionality reduction: on the one hand, downsampling, filtering and quantization are mandatory when prosodic data are sonified by means of acoustic instruments; on the other hand, a reliable mimesis (i.e. the relationship of the represented to the "essence" of its original) must be accomplished [34]. But making musical instruments 'speak' is a necessary but not a sufficient condition for sonification designers, nor for many composers. To match a successful signification of prosodic data with the artist's aesthetic interests, two hurdles can be turned into opportunities.

In the author's experience, the use of common musical instruments introduces potentially misleading artefacts based

on the subjective expectations and knowledge of the listener. But rather than causing a problem, the audio-visual cues provided by the performer tend to draw the listener's interest toward an inspection of the prosodic data. Moreover, both real acoustic instruments and musical notation resist the 'pressure' of prosody to be sonified. The tensions emerging from this resistance put stringent limits on the accuracy of the sonification. These, however, are the constraints that tie together both prosodic and musical communication. This is the very reason which may make sonifying prosody through acoustic instruments compelling for both composers and researchers.

5. REFERENCES

[1] Baird, A., Song, M., & Schuller, B, "Interaction with the Soundscape: Exploring Emotional Audio Generation for Improved Individual Wellbeing", in *International Conference on Human-Computer Interaction*. Springer, Cham, pp. 229-242, 2020.

[2] Tanenhaus, M. K., Kurumada, C., & Brown, M., "Prosody and intention recognition", in *Explicit and implicit prosody in sentence processing*. Springer, Cham, pp. 99-118, 2015.

[3] Bolinger, D., "Where does intonation belong?" *Journal of Semantics*, vol. *2*, no. 2, pp. 101-120, 1983.

[4] Thompson, W. F., & Balkwill, L. L. "Decoding speech prosody in five languages", *Semiotica*, vol. 158, pp. 407-424, 2006.

[5] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T., "Speech emotion recognition using deep learning techniques: A review". IEEE Access, vol. 7, pp. 117327-117345, 2019.

[6] Anagnostopoulos, C. N., Iliou, T., & Giannoukos, I., "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011". *Artificial Intelligence Review*, vol. 43 no. 2, pp. 155-177, 2015.

[7] J. Nichols. Kassler, J. C., "Representing speech through musical notation". *Journal of Musicological Research*, vol. 24, no. 3-4, pp. 227-239, 2005.

[8] Pearl, J., "Eavesdropping with a master: Leoš Janáček and the music of speech", *Empirical Musicology Review*, vol. 1, no, 3, 2006.

[9] Trayanova, B., & Joachim, D., "Mapping speech signals to musical scores through prosodic extraction", in *Proceedings. (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 3, pp. iii-249, 2005.

[10] Cifariello Ciardi, F., "Dalla prosodia alla musica strumentale: una sfida compositiva", in Lorenzo Cardilli, Stefano Lombardi Vallauri (ed.), *L'arte orale. Poesia, musica, performance*, Torino, Italy: Accademia University Press, 2020.

[11] Lindborg, P., "About TreeTorika: Rhetorics, CAAC and Mao", in Bresson, Agon, & Assayag (ed.), *OM Composer's Book*, *2,* 2008.

[12] Kim, D. S., "Perceptual phase redundancy in speech", in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings,* vol. 3, pp. 1383-1386), June 2000.

[13] Pellegrino, F., Coupé, C., & Marsico, E., "A cross-language perspective on speech information rate". *Language*, pp. 539-558, 2011.

[14] Dellwo, V., and Wagner, P., "Relationships between Speech Rhythm and Rate", in *Proceedings of the 15th ICPhS*, Barcelona, pp. 471-474, 2003.

[15] Zhu, Q., & Alwan, A., "On the use of variable frame rate analysis in speech recognition", in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*. vol. 3, pp. 1783-1786, June 2000.

[16] Mertens, P., "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model", in *Proceedings of Speech Prosody 2004*, Nara, Japan, pp. 23–26, 2004.

[17] Craik, F. I., "Encoding: A cognitive perspective". *Science of memory: Concepts*, 129-135. 2007.

[18] Siedenburg, K., & McAdams, S., "The role of long-term familiarity and attentional maintenance in short-term memory for timbre", *Memory*, vol. 25, no. 4, pp. 550-564. 2017.

[19] Assayag, G., Rueda, C., Laurson, M., Agon, C., & Delerue, O., "Computer-assisted composition at IRCAM: From PatchWork to OpenMusic", *Computer music journal*, vol. 23 no.3, pp. 59-72, 1999.

[20] https://github.com/openmusic-project/OM-pm2

[21] https://vimeo.com/showcase/3170584

[22] https://soundcloud.com/fabio-cifariello-ciardi/riforma-9-instruments-and-pre-recorded

[23] https://youtu.be/NuZMVha2lUs

[24] https://soundcloud.com/fabio-cifariello-ciardi/cupio-dissolvi-for-8-double-basses

[25] Wesenick, M. B., & Schiel, F., "Applying speech verification to a large data base of German to obtain a statistical survey about rules of pronunciation". In *Third International Conference on Spoken Language Processing*. 1994.

[26] Mertens, P. & Alessandro, Ch. d', "Pitch contour stylization using a tonal perception model". *Proc. Int. Congr. Phonetic Sciences,* vol. 13, no. 4, pp. 228-231, 1995.

[27] Riecke, L., van Opstal, A. J., & Formisano, E., "The auditory continuity illusion: A parametric investigation and filter model", *Perception & Psychophysics*, vol. 70, no. 1, pp. 1-12, 2008.

[28] https://soundcloud.com/fabio-cifariello-ciardi/appunti-per-amanti-simultanei

[29] Bolinger, D. L., "A theory of pitch accent in English", *Word*, vol. 14, no. 2-3, pp. 109-149, 1958.

[30] Mozziconacci, S., "Prosody and emotions", in *Speech Prosody 2002, International Conference*. 2002.

[31] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, & Hirschberg, J., "ToBI: A standard for labeling English prosody", in *Second international conference on spoken language processing*. 1992.

[32] https://obamawhitehouse.archives.gov/the-press-office/2016/01/05/remarks-president-common-sense-gun-safety-reform

[33] Vickers, P., Sonification and music, music and sonification. In *The Routledge Companion to Sounding Art* (pp. 135-144). London, UK: Taylor & Francis, 2016.

[34] Iosafat, D., "On Sonification of Place: Psychosonography and Urban Portrait", *Organised Sound,* Vol. 14, no. 1, pp. 47–55. 2009.

[35] http://sdif.sourceforge.net/standard/types-main.html

[36] https://youtu.be/UROVLQb60Pc