MAPPING SPEECH SIGNALS TO MUSICAL SCORES THROUGH PROSODIC EXTRACTION

Boyanna Trayanova, Dale Joachim

Tulane University
Department of Electrical Engineering and Computer Science
Tulane University, New Orleans, LA

ABSTRACT

Detached from its semantic content, a speech signal can be interpreted as a musical structure, containing rhythm, intonation, timbre and pitch. The musical components of speech can be extracted through algorithmic prosodic analysis to be mapped into musical notation. In this paper we present a system that extracts pitch and rhythm from an utterance for mapping into a musical score.

1. INTRODUCTION

In conversational speech, speakers continuously vary the pitch, tempo and attack of their vocal delivery. The resulting spoken utterance can be interpreted as a mixture of rhythm, intonation, timbre and pitch. Prosodic feature extraction analysis highlights these musical components of speech. In this paper we suggest a mapping of speech into musical notation as a stepping stone to speech based music composition or a novel form of speech encoding.

2. PREVIOUS WORK

Previous work in this field ranges from the extraction of prosody from a speech signal to the assignment of prosodic structure to speech and music synthesis. Buder and Eriksson [1] perform prosodic analysis on a speech signal to determine rhythmic units by using autocorrelation analysis for fundamental frequency extraction and identifying vowel onsets from the spectrographic analysis. Farinas and Pellegrino [2] propose a rhythm modeling system that segments a speech signal into pseudo-syllables through a vowel detection algorithm based on spectral analysis of the signal. Jensen and Andersen [3] introduce a feature extraction based method for real-time estimation of the beat interval in music. Jarina, O'Connor, Marlow and Murphy [4] present a method for discriminating between speech and music based on rhythm detection. Atterer [5] proposes a model that assigns prosodic structure to text for speech synthesis. Fritsch and Viccari [6] showcase a rule-based generator of musical compositions which takes into account musical parameters such as notes, durations and intensities. Hainsworth and Macleod [7] propose an algorithm that extracts and transcribes the bass line given a polyphonic audio stream. Miranda [8] designed a system that extracts the prosody of a spoken signal by computing the pitch envelope and energy contour of the signal and re-synthesizes it using these parameters and a modulator. The pitch envelope is extracted by finding periodicity in a windowed speech signal using an autocorrelation-based technique and computing the frequency of the periodic segments. The energy contour of the signal is extracted by convolving squared values of each sample in the signal with a smooth bell-shaped curve with a very low peak-side lobe. Although Miranda's module resynthesizes the speech signal, it does not perform any mapping of the speech into music.Our proposed system is based on Miranda's design, with the addition of mapping rules to convert a spoken signal into a musical score.

3. SYSTEM OVERVIEW

An overview of the system and its components is shown in Figure 1. Each component is explained in detail below.

3.1. Pitch Envelope Extraction

The pitch of a speech signal is extracted via the autocorrelation method. The autocorrelation method works by finding the periodicity of a signal through autocorrelation analysis. In order to overcome the time-varying nature of speech signals, the autocorrelation method is confined to stationary windowed segments of speech signals. The pitch extraction algorithm computes the autocorrelation of the signal in a given window of length L and, since the autocorrelation function has a global maximum at time $t = 0$, it finds a location T_0 of maximum autocorrelation that is an offset K away from the global maximum. A small time lag offset is used to bypass low frequency peaks in the autocorrelation function. The frequency of the framed signal is then $F_0 = 1/T_0$ with corresponding pitch $p(i) = F_s/F_0$ where

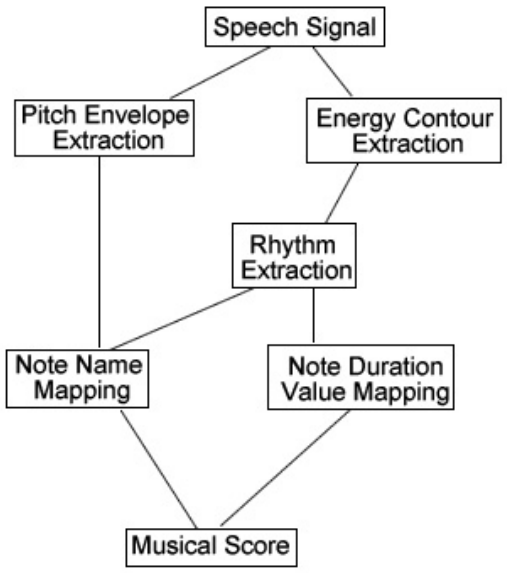

Fig. 1. Overview

$p(i)$ is the pitch of frame i in Hertz and Fs is the sampling frequency of the signal in Hertz.

3.2. Energy Contour Extraction

The energy contour of the signal is used to determine note onsets and durations in the rhythm extraction algorithm. The contour is extracted by computing an energy index for each frame k,

$$e(k) = \sum_{n=1}^{N} x(n)^2 \tag{1}$$

where $x(n)^2$ is the squared value of the signal at sample n, and N is the total number of samples in the frame. To alleviate the effects of noise, the signal is filtered of sampling values below a given threshold before the extraction of the energy contour. The filter can be defined as follows:

$$x(n) = \begin{cases} 0, & |x(n)| < \alpha \\ x(n), & |x(n)| \geq \alpha \end{cases} \tag{2}$$

where $x(n)$ is the value of the signal at sample n, and α is the threshold value. A threshold value at 10% of the maximum energy proved to be effective in our experimental signal.

3.3. Determining Rhythm From The Energy Contour

In order to produce a mapping from a speech signal into a musical score, rhythm must be extracted from the signal.

Table 1. Mapping rules for translating the duration of an utterance or pause into musical notation. R represents the relative ratio of the note duration in comparison with the longest duration in the signal.

Relative Ratio range	Note Type
$3/4 < r \leq 1$	Half
$1/2 < r \leq 3/4$	Dotted quarter
$3/8 < r \leq 1/2$	Quarter
$1/4 < r \leq 3/8$	Dotted eighth
$1/8 < r \leq 1/4$	Eighth
$r \leq 1/8$	Sixteenth

The rhythm extraction module parses the energy contour and takes into account the verbal stresses and pauses. The algorithm looks for ascending and descending trends in the energy contour. It isolates syllables and pauses by flagging changes such as descending then ascending energy and descending energy followed by energy constant at zero. The rhythm extraction module also computes the duration of the rhythmic element and notates the starting frame. The algorithm makes a distinction between pauses and utterances. These distinctions are stored in the form of flags and are subsequently used to map utterances into notes and pauses into rests.

3.4. Note Duration Value Extraction

In musical notation, note durations are represented as fractions that correspond to a ratio of beats per measure (such as quarter and eighth notes). Smaller fractions represent shorter note durations. To add the value of half of the duration of the note to the existing duration, a dot notation is used (dotted quarter notes, for example). In order to translate the duration of a syllable or a verbal pause in terms of frames into musical notation, our module uses relative ratios. The algorithm finds the longest duration of all the rhythmic units and divides all other durations by that value. The computed ratio is then used in a table lookup to perform the mapping. The table lookup is represented in Table 1. By default, the utterance or pause of the longest duration is mapped to a half note.

4. MAPPING PITCH TO NOTE

The pitch of a given note in Hertz can be mapped to a note name by using the logarithm of the frequency as an index within a table of notes and octaves. Computation of this index value is performed as follows:

$$d = \log_2(f) - \log_2\left(\frac{l_f}{1/N_o}\right) \tag{3}$$

where f is the frequency of the given note, l_f is the lowest frequency to be mapped to a note name (in this case l_f is a low A of 55 Hertz), and N_o is the number of notes per octave (N_o=12 in this case). Note names and octaves are determined as follows:

$$n_{name} = remainder\left(\frac{d}{N_o}\right) + 1 \quad (4)$$

$$n_{octave} = \lfloor \frac{(d-3)}{N_o} + 1 \rfloor \quad (5)$$

where n_{name} is the index to a table of increasing note names ranging from A to A flat, and n_{octave} is the octave number of the given note.To determine the note onsets, the starting frames of the extracted rhythmic elements are used to index the array of pitch values. Once note names and durations are generated for the entire sequence, the speech signal can be plotted on a musical staff.

5. AN EXAMPLE OF A SPEECH TO MUSIC MAPPING

For the purposes of demonstrating our speech to music mapping system, let us examine the utterance *"Hi, my name is Michelle. I go to Tulane"* spoken by an adult female speaker. The original waveform is represented in Figure 2. The extracted pitch envelope is shown in Figure 3, and the extracted energy contour is represented in Figure 4. After an analysis of the computed energy contour, our rhythm extraction module found 19 rhythmic feature elements including 5 rests and 12 notes. The durations of the rhythmic features ranged from a half note down to a sixteenth note. The extracted pitch values spanned approximately 2 octaves ranging from a high D to a low A flat. The musical score representation of the spoken utterance is shown in Figure 5.

6. CONCLUSION

In this paper, we presented a system for mapping spoken utterances into musical scores. The overall algorithm begins by extracting the pitch envelope and energy contour of the given speech signal. The energy contour is then analyzed for rhythmic features. The rhythmic features are translated into note duration values and the pitch envelope is mapped to corresponding note names and octaves.

Although the current version of the system performs a mapping of a monophonic spoken signal into a monophonic musical score, there are vast possibilities for extension and improvement. Currently, the system is unable to capture fluctuations in the pitch of note. This is particularly apparent in notes of long durations. In the language of musical notation this phenomenon is represented as a slur. We plan to extend the capabilities of our system to include the mapping of slurred pitches to the appropriate musical notation

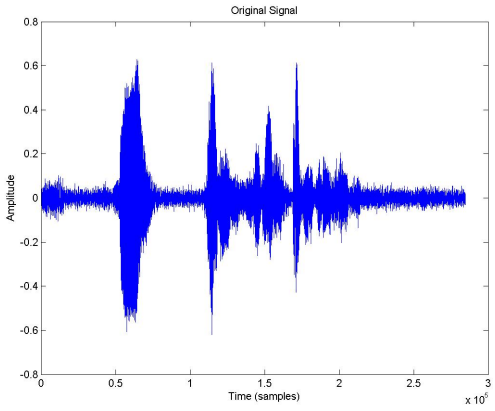

Fig. 2. Time domain representation of the spoken signal, "*Hi my name is Michelle. I go to Tulane,* sampled at 44.1 kHz.

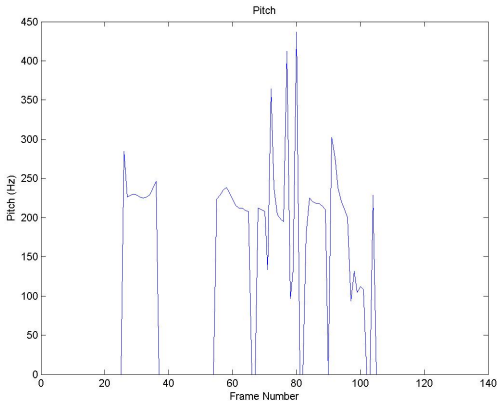

Fig. 3. The extracted pitch envelope of the speech signal of Figure 2.

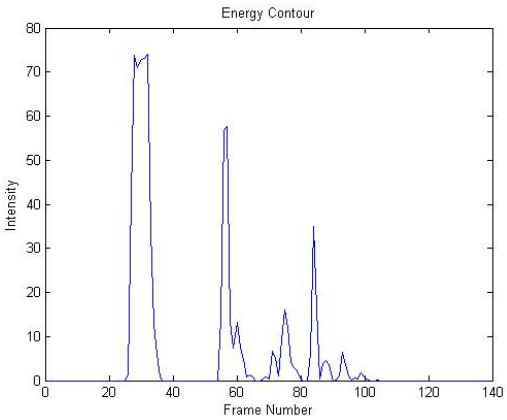

Fig. 4. Extracted energy contour of the spoken signal of Figure 2.

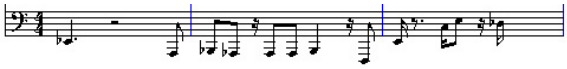

Fig. 5. Musical Score representing spoken signal of Figure 2.

thereby creating a more accurate musical representation of the sounds of the human voice. Another shortcoming of the current system is its inability to account for the amplitude (volume envelope) of the original signal in the final musical score mapping. In musical notation, the volume envelope is represented in the form of gradual dynamic markings such as crescendos and decrescendos as well as note-specific dynamic markings. We would like to extend our mapping system to be express the dynamics of the input signal and map them accordingly in the musical score. We would also like to extend our system to generate polyphonic musical scores from monophonic input signals. Although a polyphonic output will not accurately represent the sound of the human voice, this feature will have considerable compositional merit.

7. REFERENCES

[1] E.H. Buder and A. Eriksson, "Time-series analysis of conversational prosody for the identification of rhythmic units," in *Proceedings of the* 14^{th} *International Congress of Phonetic Sciences*, San Francisco, Aug 1999.

[2] J. Farinas and F. Pellegrino, "Automatic rhythm modeling for language identification," *Eurospeech '01*, vol. 4, pp. 2539–2542, 2001.

[3] K.Jensen and T.H. Andersen, "Real-time beat estimation using feature extraction," in *In Proceedings of the Computer Music Modeling and Retrieval Symposium, Lecture Notes in Computer Science*, Springer Verlag, 2003.

[4] R. Jarina, N. O'Connor, S. Marlow, and N. Murphy, "Rhythm detection for speech-music discrimination in mpeg compressed domain," in *Proc. DSP 2002 - 14th International Conference on Digital Signal Processing*, Santorini, Greece, July 2002.

[5] M. Atterer, "Assigning prosodic structure for speech synthesis: A rule-based approach. submitted to prosody," *Submitted to Prosody 2002*.

[6] E. Fritsch and R. M. Vicari, "Camm - automatic composer of musical melodies," in *XIV Congresso da SBC*, Caxambu, 1994.

[7] S.W. Hainsworth and M.D. Macleod, "Automatic bass line transcription from polyphonic music," in *In Proc. International Computer Music Conference*, Havana, Cuba, 2001.

[8] E.R. Miranda, "Computer-aided song design: Prosody as scaffolding," in *Proceedings of Annual Congress of the Brazilian Computer Science Society (SBC) - SBCM*, Universidade Federal do Ceara, July 2001.