Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Identification of Speech Transients Using Variable Frame Rate Analysis and Wavelet Packets

Daniel M. Rasetshwane, J. Robert Boston, *IEEE Member*, Ching-Chung Li, *IEEE Fellow*
Department of Electrical and Computer Engineering, University of Pittsburgh,
Pittsburgh, PA 15261, USA

Abstract -- **Speech transients are important cues for identifying and discriminating speech sounds. Yoo *et al* and Tantibundhit *et al* were successful in identifying speech transients and, emphasizing them, improving the intelligibility of speech in noise [3] [4]. However, their methods are computationally intensive and unsuitable for real-time applications. This paper presents a method to identify and emphasize speech transients that combines subband decomposition by the wavelet packet transform with variable frame rate (VFR) analysis and unvoiced consonant detection. The VFR analysis is applied to each wavelet packet to define a transitivity function that describes the extent to which the wavelet coefficients of that packet are changing. Unvoiced consonant detection is used to identify unvoiced consonant intervals and the transitivity function is amplified during these intervals. The wavelet coefficients are multiplied by the transitivity function for that packet, amplifying the coefficients localized at times when they are changing and attenuating coefficients at times when they are steady. Inverse transform of the modified wavelet packet coefficients produces a signal corresponding to speech transients similar to the transients identified by Yoo *et al* and Tantibundhit *et al*. A preliminary implementation of the algorithm runs more efficiently.**

I. INTRODUCTION

Several studies have shown that speech transients are important cues for identifying and discriminating speech sounds [1] [2]. Speech transients are associated with consonants, transitions between consonants and vowels, and transitions within some vowels. Yoo *et al* and Tantibundhit *et al* identified speech transients and showed that selective amplification of speech transients improves the intelligibility of speech in noise. Yoo used special time-varying bandpass filters to decompose highpass filtered speech into quasi-steady-state and transient speech components [3]. They modified speech to make it more intelligible by combining the amplified transient component with the original speech and adjusting the energy of the modified speech to be equal to that of the original speech. Using a psychoacoustic procedure (modified rhyme test), they showed that the modified speech has higher word recognition scores at low signal-to-noise ratios (SNR) than the original speech.[1]

Tantibundhit modified an algorithm described by Daudet and Torresani and used it to decompose speech into tonal, transient and residual components [4] [5]. In their modification, they used hidden Markov chain models and hidden Markov tree models to describe statistical dependencies in modified cosine transform coefficients and wavelet transform coefficients, respectively. A tonal speech component was obtained by the inverse of the significant modified cosine transform coefficients, and the transient component was obtained by the inverse of the significant wavelet transform coefficients. Tantibundhit *et al* also showed that speech modified with the transient component provides improved word recognition rates at low SNR.

The algorithms of Yoo *et al* and Tantibundhit *et al* were successful in identifying speech transients. However, because of their complexity, these algorithms are computationally intense and have not been implemented to run in real time. We propose a wavelet-packet based method for identifying a transient speech signal that emphasizes the same parts of speech as the transient speech components of Yoo and Tantibundhit. Our algorithm is much more efficient than their algorithms. The algorithm combines wavelet-based subband decomposition with variable frame rate (VFR) analysis and unvoiced consonant detection.

Variable frame rate processing of speech has been used by several researchers to reduce the amount of data processed in automated speech recognition systems and to improve the performance of these systems [6] [7] [8]. VFR analysis varies the window step size dynamically, using a small window step size when the speech signal is changing rapidly (retaining most of the frames) and a large window step size when the speech signal is changing slowly (discarding most of the frames). A consequence of this approach is the identification of time intervals during which the characteristics of the speech signal are changing rapidly.

In VFR implementations, a function that describes the rate of change in the speech signal is compared to a threshold to decide whether a given frame should be kept or discarded. This function, which we call the transitivity function, is large and positive when spectral characteristics of the speech are changing rapidly and near zero when spectral characteristics are changing slowly. We use this function to extract the transient speech signal.

Since speech is a multi-component signal with time-varying frequency and amplitude, transitions may occur at different times in different frequency bands. To capture these effects, the subband decomposition by wavelet packets is used, and a separate transitivity function for each frequency band (packet) is computed, allowing separate emphasis of speech transients in each packet.

The transitivity function has larger peaks for transitions into and out of high energy formants than for transitions associated with low energy consonants such as unvoiced consonants. To increase the peaks of the transitivity function

[1] This research was supported in part by the Office of Naval Research under grant number N000140310277.

that correspond to unvoiced consonants, simple unvoiced consonant detection that is based on the average short-time zero-crossing rate (ZCR) and the average short-time energy (STE) is utilized. Unvoiced speech has high ZCR and low STE [9]. Since all vowels are voiced, the ZCR and the STE can be used to identify unvoiced consonants.

Section 2 of the paper describes the proposed method. In Section 3, the importance of combining VFR with the wavelet-packet based subband decomposition is illustrated using a synthetic signal. Then the transient speech signal obtained using our method is compared to the transient speech components obtained by Yoo *et al* and Tantibundhit *et al*. Finally, the findings are summarized and discussed in Section 4.

II. THE PROPOSED METHOD

The method uses the wavelet packet transform with Daubechies-8 mother wavelet, which is a relatively symmetric orthogonal wavelet [10]. The speech signal, originally sampled at 11025 Hz, is decomposed to a depth of 5, resulting in 32 packets, each with a sequence of wavelet coefficients. We stopped at the scale level 5 because at higher decomposition levels, the number of wavelet coefficients in a packet are too few to allow for a reliable computation of the transitivity function.

For each packet, a VFR technique proposed by Le Cerf *et al* is used to compute a transitivity function that characterizes the transition rate of the wavelet coefficients of that packet. The wavelet coefficients are framed using a 16 ms. Hamming window and a window step size of 8 ms. For each frame, 12 Mel-frequency cepstral coefficients (MFCC) are computed from the wavelet packet coefficients (not from the time domain signal) to model the characteristics of these wavelet coefficients. The first derivatives of the MFCC over the frame sequence are computed and smoothed using a lowpass filter. The Euclidean norm of the derivatives is defined as the transitivity function. The 16 ms. window size was found to provide a good balance between time and frequency resolution and result in a transitivity function that is smooth but sufficiently sensitive to changes in spectral characteristics.

The number of samples in the transitivity function is equal to the number of frames of the wavelet coefficients it is computed from. To enable direct multiplication of the wavelet coefficients by the transitivity function, the transitivity function is linearly interpolated to have as many samples as the wavelet coefficients.

The transitivity function computed for a wavelet packet is large when the wavelet coefficients of that packet are changing rapidly and is small when the wavelet coefficients are changing slowly. To deemphasize the regions when the transitivity function is small, a threshold is applied to the transitivity function as follows. A sample of the computed transitivity is set to zero when its value is less than the threshold which is chosen as the mean value of the transitivity function of that packet. Abrupt changes from non-zero-valued samples to zero-valued samples of the transitivity function are smoothed by replacing the five zero-valued samples of the transitivity function following and preceding a non-zero-valued sample by a half period of a cosine function. Samples of the transitivity function that are greater than the threshold are not changed.

To determine unvoiced consonant intervals in speech, the STE and ZCR of the speech signal are computed using a 40 ms. window and a window step size of 1 sample. Here, a long window size (longer than 16 ms) is used so that the STE and ZCR are smooth. A segment of speech is considered to be an unvoiced consonant if, for that segment, the STE is less than the mean STE for the word and the ZCR is greater than the mean ZCR for the word. The transitivity function is amplified during unvoiced consonants intervals to have a value that equal the maximum value of the transitivity function for that packet.

The wavelet coefficients for each packet of the wavelet decomposition are then multiplied by the transitivity function obtained for that packet. Wavelet coefficients that occur when the transitivity function is large are amplified and wavelet coefficients that occur when the transitivity function is small are attenuated or set to zero. Finally, the inverse wavelet packet transform of the modified coefficients is computed, giving a signal that we call the transient speech signal.

To demonstrate the value of subband decompositions, the algorithm is applied to a synthetic signal consisting of a steady tone and a chirp, and transient signals for the synthetic signal obtained with and without using the wavelet packet transform are compared.

To extract a transient speech signal without using wavelet packets, a transitivity function is computed directly from the time-domain speech signal. The speech signal is framed using a 16 ms. Hamming window and a window step size of 8 ms. For each frame, 12 Mel-frequency cepstral coefficients are computed to model the speech signal. Then the first derivatives of the MFCC are computed and smoothed. The Euclidean norm of the first derivatives of the MFCC gives the transitivity function. Finally, values of the transitivity function below the mean value of the transitivity function are set to zero, and the transitivity function is amplified during unvoiced consonant intervals, as determined by the ZCR and the STE. The speech signal is then multiplied directly by the transitivity function.

III. EXPERIMENTS

Tests were carried out on a synthetic signal to demonstrate the importance of subband decomposition by wavelet packets. The synthetic signal consisted of transient activities which model abrupt changes in amplitude and frequency occurring at different times in different frequency bands. The synthetic signal, whose narrowband spectrogram is shown in Fig. 1, has two components. One component is a quasi-steady-state tone and the second component, starting 50 ms after the first, includes a quasi-steady-state tone and a transition, via a chirp, to another quasi-steady-state tone of higher frequency. The spectrogram was obtained using a Hanning window of length 10 ms. and a window overlap of 1 ms. For this synthetic signal, transient activities modeling changes in amplitude are the onsets and offsets of the two components, and transient activity modeling changes in frequency is the chirp of the second component. The algorithm should emphasize these transitions and deemphasize the quasi-steady-state tones.

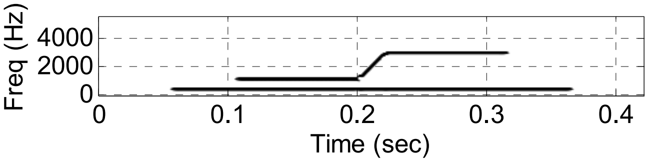

Figure 1: Spectrogram of synthetic signal.

Figure 2 shows the transient signals for the synthetic signal of Fig. 1 extracted using VFR alone (top) and VFR with subband decomposition (bottom). For the synthetic signal, unvoiced consonant detection was not used to modify the transitivity function. The transient signal obtained using VFR alone, as shown in the top part of Fig. 2, emphasizes the onsets and offsets of the two components and the chirp. However, it also includes quasi-steady-state portions of the steady component around 0.1 and 0.3 s. and between 0.15 and 0.25 s. This quasi-steady-state activity is included in the transient signal because it occurs during periods when the chirp component has transitions.

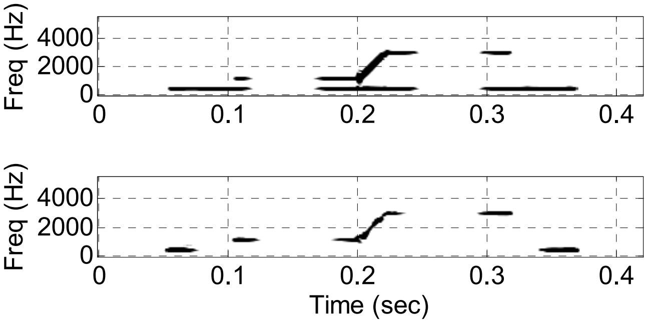

Figure 2: Demonstration of the importance of subband decomposition via a synthetic signal shown in Fig 1. Transient signals extracted using VFR alone (top) and VFR with subband decomposition (bottom).

The bottom part of Fig. 2 shows the transient signal obtained using VFR with subband decomposition. The extracted transient signal emphasizes the onsets and offsets of the two components and the chirp and de-emphasizes the steady-state portions of all three tones. With the subband decomposition provided by wavelet-packets, quasi-steady-state activity of the first component occurring while the second component has transitions is not extracted in the transient signal.

Our method has been applied to a wide range of speech material, including monosyllable consonant-vowel-consonant (CVC) and two-syllable words, and the transient extracted is similar to that obtained by Yoo *et al* and Tantibundhit *et al*. The results obtained for the word ‘bat’ spoken by a male are representative of these tests. ‘Bat’, phonetically transcribed as /bæt/, includes a voiced bilabial plosive consonant /b/, a vowel /æ/, and an unvoiced alveolar plosive consonant /t/. The time-domain waveform and wideband spectrogram for this word are shown in Fig. 3. This spectrogram and spectrograms shown later were obtained using a Hanning window of length 5 ms. and a window overlap of 1 ms.

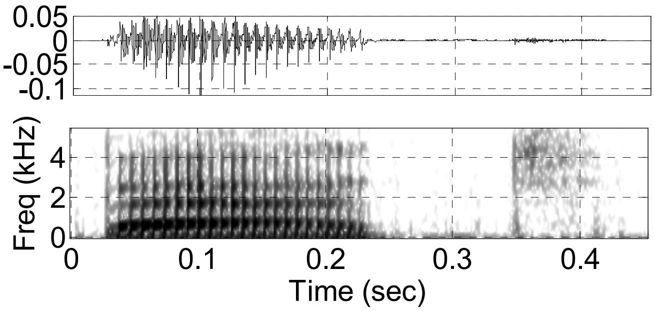

Figure 3: Time-domain waveform (top) and wideband spectrogram (bottom) for the word ‘bat’ /bæt/ spoken by a male.

The transient speech signal for the word ‘bat’ /bæt/ obtained using the proposed method is compared to the transient speech components obtained by Yoo’s method and Tantibundhit’s method in Fig. 4. The arrows in the spectrograms indicate the quasi-steady-state portions of formant activity that have been removed. In all three transient signals, the quasi-steady-state portions of the first and second formants, indicated by the two lower arrows in the spectrograms, are removed, and the beginning and ending of these formants are emphasized. The alveolar plosive consonant /t/, from 0.35 s., is also emphasized. In addition to removing quasi-steady-state activity of the first and second formants, our transient speech signal, like the transient speech component of Tantibundhit *et al*, also removed quasi-steady-state portions of the third and fourth formants, indicated by the two upper arrows in the spectrograms, and emphasized the beginning and ending of these formants.

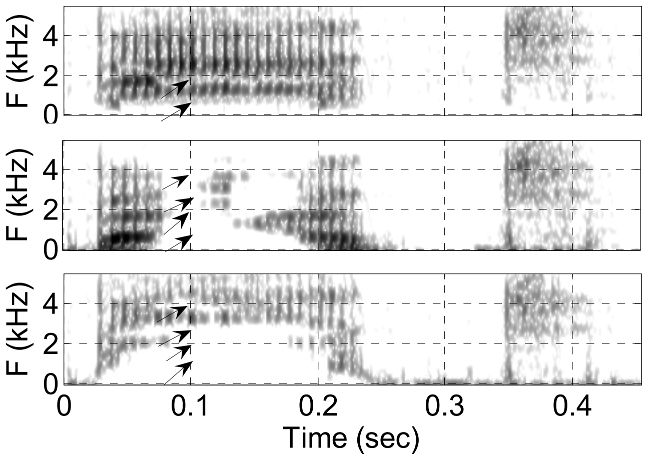

Figure 4: Transient speech signals for the word ‘bat’ /bæt/ obtained using Yoo *et al*’s method (top), our method (middle) and Tantibundhit *et al*’s method (bottom).

The transient signal obtained by Yoo *et al* was the most intelligible and the easiest to identify as the word 'bat', whereas Tantibundhit *et al*'s transient signal was the least intelligible and was not only soft but also not easily identified as the word 'bat'. The intelligibility of our transient signal was between the intelligibilities of the other two transient signals but could be easily identified as the word 'bat'. The transient speech component of Tantibundhit included the lowest energy of the original speech, 0.7 %. The transient speech component of Yoo included 7.7 % of the energy of the original speech, whereas our transient speech signal included 12.7 %. However, highpass filtering performed by Yoo alone removed about 70 % of the energy of the original speech.

Table I shows the energy in the transient speech signals averaged over 10 CVC words, as a percentage of the energy of the original speech, obtained using the methods of Yoo, Tantibundhit, and our method. Tantibundhit method has the lowest energy and our method has the highest energy, about twice the energy of Yoo.

Table I: Average energy in the transient speech signals.

Yoo *et al*'s method	5.14 %
Tantibundhit *et al*'s method	0.74 %
Proposed method	11.07 %

The three methods were implemented in MATLAB (The MathWorks, Inc). The average computation time for Yoo's algorithm was approximately 50 times the duration of the speech processed and for Tantibundhit's algorithm was approximately 120 times the duration of the speech processed. The computation time of our algorithm was about 5 times the duration of the speech processed.

IV. DISCUSSION AND CONCLUSION

A method that identifies speech transients is presented. The algorithm uses subband decomposition based on wavelet packets, variable frame rate analysis and unvoiced consonant detection at each packet to identify when speech transitions occur. Subband decomposition allows the detection of transients occurring at different times in different frequency bands and reduces the amount of quasi-steady-state activity that would be identified as transient.

The transient speech signal identified using our method emphasizes the same parts of the original speech as the transient speech components of Yoo *et al* and Tantibundhit *et al*. Spectrograms of Fig. 4 show that, like the transient components of Yoo and Tantibundhit, our transient signal removes quasi-steady-state formant activity and emphasizes the beginning and ending of formants. Further comparison of our transient speech signal to the transient speech component of Tantibundhit shows that the former emphasizes the beginning and ending of formants more strongly.

The transient speech component extracted by Yoo preserves formant activity more than those obtained by the other two methods because it retains the third and fourth formants. As a result, it was the most intelligible of the three transient signals. Table I shows that, for the set of words processed, the energy of the original speech signal included in our transient speech signal is twice the energy included in the transient speech component of Yoo. However Yoo used highpass filtering, which removed approximately 70 % of the speech energy, including low frequency transients. Our method retained these transients.

The transient speech signals obtained with the proposed algorithm emphasizes the same parts of speech as the transient speech components of the algorithms proposed by Yoo and Tantibundhit. Their algorithms are computationally intensive and unsuitable for real-time applications. Our method identifies speech transients much more efficiently than their methods and requires a much lower computation time. We believe our method can be implemented to run in real-time.

REFERENCES

[1] Liberman, A. M., Delattre P. C., Cooper, F. S., Gerstman, L. J., "The Role of Consonant-Vowel Transitions in the Perception of the Stop and Nasal Consonants", *Psychology Monographs: General and Applied*, vol. 68, no. 8, pp. 1-13, 1954.

[2] Fant, G., *Speech Sounds and Features*, MIT Press, Cambridge, MA, 1973.

[3] Yoo, S., Boston, J.R., Durrant, J.D., Kovacyk, K., Karn, S., Shaiman, S. E-Jaroudi, A. & Li., C.C., "Relative energy and intelligibility of transient speech components", *Proceedings of IEEE ICASSP '05,* vol. 1, pp. 69-72, March 2005.

[4] Tantibundhit, C., Boston, J.R, Li, C.C. & El-Jaroudi, A., "Automatic Speech Decomposition and Speech Coding Using MDCT-based Hidden Markov Chain and Wavelet-based Hidden Markov Tree Models", *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 207 – 210, October 2005.

[5] Daudet, L. & Torresani, B. (2002), "Hybrid representation for audiophonic signal encoding", *Signal Processing,* vol. 82, pp. 1595 – 1617, 2002.

[6] Le Cerf, P. & Van Compernolle, D., "A New Variable Frame Rate Analysis for Speech Recognition", *IEEE Signal Processing Letters,* vol. 1, no. 12, pp. 185-187, December 1994.

[7] Ponting, K. M. & Peeling, S. M., "The Use of Variable Frame Rate Analysis in Speech Recognition", *Computer Speech and Language,* vol. 5, pp. 169-179, April 1991.

[8] Zhu, Q. & Alwan, A., "On the Use of Variable Frame Rate Analysis in Speech Recognition", *Proceedings of IEEE ICASSP '00,* vol. 3, pp. 1783-1786, June 2000.

[9] Rabiner, L.R. & Schafer, R.W., *Digital Processing of Speech Signals*, Prentice Hall, Upper Saddle River, NJ, 1978.

[10] Daubechies, I., *Ten Lectures on Wavelets,* Philadelphia: SIAM, CBMS-NSF Regional Conference in Applied Mathematics 61, 1992.