SPEECH ENHANCEMENT USING TRANSIENT SPEECH COMPONENTS

C. Tantibundhit[1], *J. R. Boston*[1,2], *C. C. Li*[1], *J. D. Durrant*[2], *S. Shaiman*[2], *K. Kovacyk*[2], *A. El-Jaroudi*[1]

Department of Electrical and Computer Engineering[1] and
Department of Communication Science and Disorders[2]
University of Pittsburgh, Pittsburgh, PA 15261, USA

ABSTRACT

This paper describes an algorithm to decompose speech into tonal, transient, and residual components. The algorithm uses an MDCT-based hidden Markov chain model to isolate the tonal component and a wavelet-based hidden Markov tree model to isolate the transient component. We suggest that the auditory system, like the visual system, is probably sensitive to abrupt stimulus changes and that the transient component in speech may be particularly critical to speech perception. To test this suggestion, the transient component isolated by our algorithm was selectively amplified and recombined with the original speech to generate enhanced speech, with energy adjusted to be equal to the energy of the original speech. The intelligibility of the original and enhanced speech was evaluated in eleven human subjects by the modified rhyme protocol. The word recognition rates show that the enhanced speech can provide substantial improvement in speech intelligibility at low SNR levels (8% at –15 dB, 14% at –20dB, and 18% at –25 dB).

1. INTRODUCTION

The auditory system, like the visual system, may be sensitive to abrupt stimulus changes, and the transient component in speech may be particularly critical to speech perception. If this component can be identified and selectively amplified, improved speech perception in background noise may be possible. Yoo *et al.* employed a formant-tracking filter to remove the dominant formant energy from speech and they investigated the use of this component to enhance speech in noise [1]. However, the resulting transient retained a significant amount of energy during what would appear to be tonal regions of the speech. In this paper, we define an alternative method to identify a transient component. The algorithm decomposes speech into three components, based on the approach of Daudet and Torrésani [2], as signal = tonal + transient + residual components. The modified discrete cosine transform (MDCT), which provides good estimates of a locally stationary signal, was utilized to estimate the tonal component. The wavelet transform, which provides good results in encoding signals exhibiting abrupt temporal changes, was applied to estimate the transient component.

Daudet and Torrésani were interested in improved speech coding, and they identified tonal and transient components using the inverse transform of a small number of most significant coefficients of the MDCT and the wavelet transform, where the significant MDCT and wavelet coefficients were determined by thresholds. They assumed that the MDCT coefficients and the wavelet coefficients were independent. However, both the MDCT coefficients and the wavelet coefficients can be expected to show statistical dependencies, namely clustering and persistence properties.

Crouse *et al.* developed a probabilistic model to capture clustering and persistence properties of the wavelet transform coefficients [3] using a hidden Markov tree (HMT) model to describe the statistical dependencies of the wavelet coefficients along and across scale. They modeled the wavelet coefficients by a two-state, zero-mean Gaussian mixture, where "large" states and "small" states were associated with large variance and small variance, zero-mean Gaussian distributions, respectively. They also introduced an upward-downward algorithm for training the model.

Molla and Torrésani applied the HMT model of [3] to estimate the transient component in a musical signal [4]. They associated the transient state with a large-variance Gaussian distribution and the residual state with a small-variance Gaussian distribution. They used the statistical inference method [5], which is more robust to the numerical underflow problem than the upward-downward algorithm used in [3]. Daudet *et al.* proposed another probabilistic model to estimate the tonal component in a musical signal [6]. They applied a hidden Markov chain (HMC) model to describe the statistical dependencies of the MDCT coefficients in each frequency index, modeling the MDCT coefficients as a two-state, zero-mean Gaussian mixture. A tonal state was associated with a large-variance Gaussian distribution, and a non-tonal state was associated with a small-variance Gaussian distribution.

Our algorithm [7] is a modification of [2] that avoids using thresholds and can capture statistical dependencies between the MDCT coefficients and the wavelet coefficients by utilizing the HMC model [6] and the HMT model [4], respectively. The Viterbi algorithm [8] and the Maximum *a Posteriori* (MAP) algorithm [5], used to find the optimal state distribution, are applied to determine the significant MDCT and wavelet coefficients. The algorithm to decompose the speech signal into different components is briefly reviewed in Section 2. After identification, the transient component is selectively amplified and recombined with the original speech to generate enhanced speech. The intelligibility of the original speech and enhanced speech is evaluated by a modified rhyme test, using the protocol described in Section 3. The results are presented in Section 4, and their implications are discussed in Section 5.

2. SPEECH DECOMPOSITION AND SPEECH ENHANCEMENT

The algorithm decomposes speech into three different components as signal = tonal + transient + residual. The tonal and transient components are identified using a small number of coefficients of the MDCT and the wavelet transform, respectively. Instead of running the algorithm once, we run the algorithm twice, based on alternate

This work is supported by the Office of Naval Research under the grant number N000140310277.

projection [9], because we found that the residual component from the first iteration appeared to retain significant tonal and transient information.

2.1. Tonal Estimation

The original speech signal, $x_{\mathrm{orig}}(t)$, sampled at 11.025 kHz, was expanded by the MDCT. The half window length was set to 2.90 ms, equivalent to 32 coefficients. We found this length to be short enough that the tonal component in each time frame can be reasonably assumed to be a locally stationary signal and long enough to ensure sufficient frequency resolution. This window length also reduced the pre-echo effect [10]. The windows were half overlapped to optimize frequency localization [2]. The MDCT coefficients in each frequency index were applied to the HMC model, which is a two-state mixture of two univariate Gaussian distributions. Each MDCT coefficient was conditioned by one of two hidden states, representing tonal and non-tonal states. The tonal state was associated with the large-variance Gaussian distribution, and the non-tonal state was associated with the small-variance Gaussian distribution.

The initial parameters (weights, means, and variances) of the mixture of two univariate Gaussian distributions in each frequency index were estimated by the greedy EM algorithm [11]. The forward-backward algorithm [8] was used to train the MDCT coefficients in each frequency index until the local optimum corresponding to the maximum likelihood was reached. Then, the Viterbi algorithm [8] was used to find the optimal state distribution in each frequency index such that each MDCT coefficient was conditioned by either a tonal or a non-tonal hidden state. All of the MDCT coefficients with tonal hidden states were retained and those with non-tonal hidden states were set to zero, providing an identification of the MDCT coefficients to construct the tonal component. The tonal component, $x_{\mathrm{tone}}(t)$, was calculated by the inverse transform of those MDCT coefficients, and the non-tonal component, $x_{\mathrm{nont}}(t)$, was calculated by subtracting the tonal component from the original signal, $x_{\mathrm{nont}}(t) = x_{\mathrm{orig}}(t) - x_{\mathrm{tone}}(t)$. This method to identify the tonal component does not require a threshold.

2.2. Transient Estimation

The non-tonal component (of length N) was expanded by the wavelet transform, using the Daubechies-8, the most nearly symmetric wavelet [3]. The transform was limited to level (L = 7), resulting in $\mathrm{K} = \mathrm{N}2^{-\mathrm{L}}$ trees, where each tree was 11.61 ms long and corresponded to 128 coefficients.

The wavelet coefficients at each scale of each tree were applied to the HMT model, which is a two-state mixture of two univariate Gaussian distributions. Because there are small numbers of wavelet coefficients in each scale of each tree, especially in the coarse scale, the initial parameters of the mixture were calculated by applying the greedy EM algorithm to all wavelet coefficients in that tree.

Each wavelet coefficient was conditioned by one of two hidden states, representing a transient and a residual state. The transient state was associated with a large-variance distribution, and the residual state was associated with a small-variance distribution. Each hidden state models a random process defined by a coarse-to-fine hidden Markov tree with a constraint. The constraint is that a transition from the residual state to the transient state is not allowed ($P\{S_{\mathrm{child}} = \mathrm{Transient} | S_{\mathrm{parent}} = \mathrm{Residual}\} = 0$) [4].

The conditional upward-downward algorithm [5] was used to train the wavelet coefficients in each scale of each tree until the local optimum corresponding to the maximum likelihood was reached. Then, the parameters were tied with the parameters in the corresponding scale of its neighboring left and right trees, using robust via tying [3].

The MAP algorithm [5] was applied to find the optimal state distribution of each tree such that each wavelet coefficient was conditioned by either a transient or residual hidden state. All of the wavelet coefficients conditioned by transient hidden states were retained. Those with residual hidden states were set to zero. The transient component, $x_{\mathrm{tran}}(t)$, was obtained as the inverse wavelet transform of the retained wavelet coefficients. The residual component, $x_{\mathrm{resi}}(t)$, was calculated by subtracting the transient component from the non-tonal component, $x_{\mathrm{resi}}(t) = x_{\mathrm{nont}}(t) - x_{\mathrm{tran}}(t)$.

For the second iteration, the residual component from the first iteration was used in place of the original speech signal, and the algorithm was repeated. The resulting tonal and transient components are the summation of the tonal and the transient components from the first and the second iterations, respectively. The resulting residual component is the residual component from the second iteration.

2.3. Speech Decomposition Results

Figure 1 illustrates speech decomposition results for the mono-syllablic consonant-vowel-consonant (CVC) word "got", spoken by a male speaker. This word can be transcribed phonetically as /gɑt/. It includes a voiced velar plosive stop consonant /g/, a vowel /ɑ/, and a voiceless alveolar plosive stop consonant /t/. The spectrogram of the word is illustrated on top of the figure.

The tonal component, illustrated in the middle of the figure, includes most (99%) of the energy of the speech signal. This component predominantly includes constant frequency information of the first, second, third, and forth formant frequency. It also includes consonant hubs of the /t/ release, that appear in the high frequency range around 4-5 kHz from 0.42 to 0.45 sec.

The transient component, illustrated at the bottom of the figure, includes 1% of the total energy. It includes the /g/ release and the start of /t/ release, shown as the vertical ridges in the spectrogram at approximately 0.02 sec and 0.42 sec, respectively. It also includes most of the /t/ release, which appears as a noise pattern in the high frequency range from 0.42 sec to the end of the word. In addition, it nicely includes formant transitions from the /g/ release into the first, second, third, and the forth formants of the vowel /ɑ/ as well as transitions around the end of the vowel. For this word, the residual component was very small including approximately 0.001% of the total energy.

2.4. Speech Enhancement

Enhanced speech was generated by $x_{\mathrm{enha}}(t) = a(x_{\mathrm{orig}}(t) + b \cdot x_{\mathrm{tran}}(t))$, where x_{enha}, x_{orig}, and x_{tran} represent the enhanced, original, and transient speech, respectively. a is a factor to adjust the energy of the original and the enhanced speech to be the same. b is the transient amplification factor, chosen to be 12, based on a preliminary evaluation from factors 1 to 15. Smaller factors had little effect on the words, and larger factors introduced distortion.

Figure 2 illustrates the effects of the enhancement process on the word "got" /gɑt/ in the time domain. The enhanced speech shows more prominent /g/ release, transitions from the /g/ release into and out of the vowel formants, and the beginning and the release /t/ than the original speech.

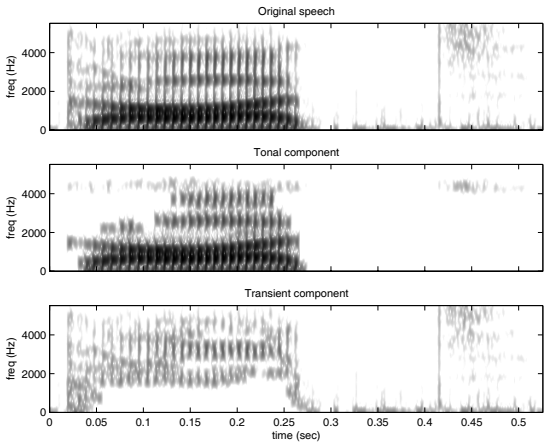

Fig. 1. Speech decomposition of "got"

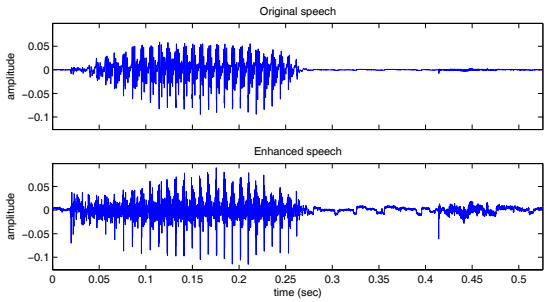

Fig. 2. Enhanced version of "got"

SNR	Mean difference	SD difference	95% CI difference	p-value
–25 dB	17.50	12.93	8.77 ∼ 26.14	0.0012
–20 dB	13.82	20.89	–0.22 ∼ 27.85	0.0530
–15 dB	7.64	11.24	0.09 ∼ 15.19	0.0479
–10 dB	3.64	15.95	–7.08 ∼ 14.35	0.4669
–5 dB	–0.73	14.51	–10.48 ∼ 9.02	0.8713
0 dB	–2.55	14.56	–12.33 ∼ 7.24	0.5749

Table 1. Differences (enhanced speech – original speech) of means, standard deviations (SDs), 95% confidence intervals (CIs), and p-values of word recognition scores

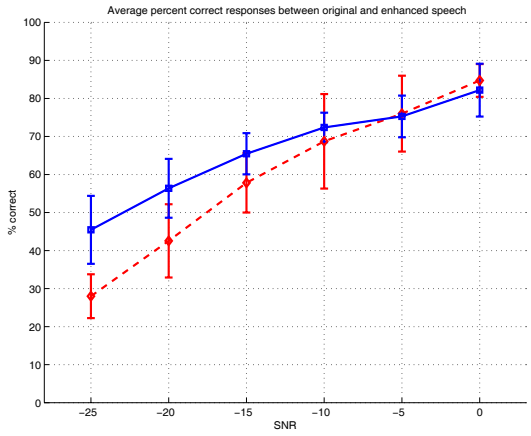

Fig. 3. Average percent correct responses between original (dashed line) and enhanced speech (solid line)

3. PSYCHOACOUSTIC TEST METHODS

Speech material was decomposed into components using the algorithm described in the previous section, and the transient component of each word was used for enhancement. The modified rhyme protocol of [12], developed from [13] and [14], was used to compare the intelligibility of enhanced speech to original speech.

The protocol was performed on eleven volunteer subjects with negative otologic histories and having hearing sensitivity of 15 dB HL or better by conventional audiometry (250 - 8 kHz). Fifty sets of rhyming monosyllabic CVC words (6 words per set for a total of 300 words), were recorded by a male speaker [13]. Among them, 25 sets differ in their initial consonants and 25 sets differ in their final consonants. Subjects sat in the sound-attenuated booth and were asked to identify a target word from a list of six words. The target word appeared on the computer screen and remained until all of six words were presented. These six words were presented at one of six SNR levels (–25, –20, –15, –10, –5, and 0 dB) using speech-weighted background noise through the right headphone. The subjects were asked to click the mouse as soon as they thought that they heard the target word. The subjects could not change an answer and could not select a previous word. The subjects were monitored during the test by skilled examiners under supervision of a certified clinical audiologist, and all subject responses were saved on the computer.

The modified rhyme protocol was composed of a training and the main test sessions. The training session made the subjects familiar with the test. The main session included 300 trials — 150 trials of the original speech and 150 trials of the enhanced speech. The 150 trials of the original and enhanced speech were equally distributed over the 6 SNR levels, giving 25 trials of original speech and 25 trials of enhanced speech at each level of background noise. The target words were randomly chosen from the 300 rhyming words. Once a chosen target word was presented, it was removed from future selections such that the same word did not occur as a target more than once.

4. RESULTS

The average percent correct responses were calculated by the subject's correct responses divided by the total numbers of stimuli. Means, standard deviations (SDs), 95% confidence intervals (CIs), and p-values of the paired-sample difference at each SNR level are summarized in Table 1. The results suggest that there are substantial difference in speech perception between the original and enhanced speech at –25dB, –20dB, and –15dB with mean differences 17.50%, 13.82%, and 7.64%, respectively. The CI differences do not include zero at –25dB (p-value = 0.0012) and at –15dB (p-value = 0.0479). The CI difference at –20dB includes zero (p-value = 0.0530), which probably occurred because of high variations in subjects.

Figure 3 shows the percent correct responses averaged for each speech type between original (dashed line) and enhanced speech

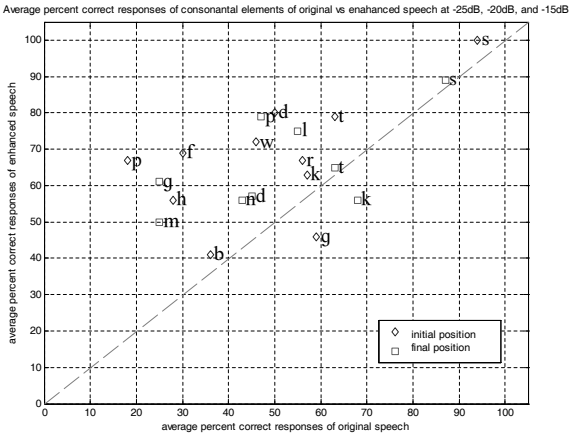

Fig. 4. Average percent correct responses according to phonetic elements in initial (◇) and final (□) positions between original and enhanced speech

(solid line) with group 95% CIs. The average percent correct responses of the original and enhanced speech increase with increasing SNR levels, and the advantage provided by enhancement decreases.

Confusions of consonantal elements in the initial and final positions were also analyzed. The motivation of this analysis is to reveal the degree of improvement in identifying various consonantal elements of the enhanced speech compared with the original speech. Because the 300 rhyming words are not phonetically balanced [13], only the initial and final consonants with high frequency of occurrences, ie more than or equal to 20, were used in this analysis.

Figure 4 illustrates the average percent correct responses of consonantal elements in the initial (11 consonants) and in the final positions (9 consonants) of original speech and of enhanced speech at –25dB, –20dB, and –15dB. These values were calculated by the numbers of correct responses divided by the total number of responses. Data points above the 45° line indicate elements that were recognized better in enhanced speech, and data points below the line indicate elements that were recognized better in original speech. Only 1 consonantal element in initial position (/g/) and 1 consonantal element in final position (/k/) were recognized less successfully in enhanced speech than in original speech. These are both plosive consonants.

5. DISCUSSION

We introduced a method to identify the transient information in speech using MDCT-based hidden Markov chain and wavelet-based hidden Markov tree models. The perception of the enhanced speech in noise is better than that of the original speech for most of SNR levels (–25, –20, –15, and –10 dB). These results suggest that the transient component is important in speech perception and emphasis of this component may provide an approach to enhance intelligibility of speech signal, especially in noisy environment.

The confusion analysis suggests that most consonants are consistently more intelligible in enhanced speech. Two plosive consonants were exceptions. The 300 rhyming words are not phonetically balanced [13] and the modified rhyme protocol [12], based on a word-monitoring task [14], does not force the subjects to make a response to every stimulus. More subjects would be required in order to analyze confusions effectively. This analysis of confusions was presented as a preliminary study to reveal the degree of improvement in identifying various consonantal elements of the enhanced speech compared with the original speech. A more complete analysis might suggest improvement in the algorithm.

6. REFERENCES

[1] S. Yoo, J. R. Boston, J. D. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C. C. Li, "Relative energy and intelligibility of transient speech information," in *Proc. IEEE Int. Conf. Acoustic, Speech, and Signal Processing*, Mar. 2005, pp. 69–72.

[2] L. Daudet and B. Torrésani, "Hybrid representation for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, 2002.

[3] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[4] S. Molla and B. Torrésani, "Hidden Markov tree of wavelet coefficients for transient detection in audiophonic signals," in *Proceedings of the Conference Self-Similarity and Applications, Clermont-Ferrand*, 2002.

[5] J. B. Durand and P. Gonçalvès, "Statistical inference for hidden Markov tree models and application to wavelet trees," Tech. Rep. 4248, Institut National de Recherche en Informatique et en Automatique, Sept. 2001.

[6] L. Daudet, S. Molla, and B. Torrésani, "Towards a hybrid audio coder," in *Proc. of the International Conference Wavelet Analysis and Applications, Chongqing, China, Jian Ping Li Editor, World Scientific*, 2004, pp. 12–21.

[7] C. Tantibundhit, J. R. Boston, C. C. Li, and A. El-Jaroudi, "Automatic speech decomposition and speech coding using mdct-based hidden Markov chain and wavelet-based hidden Markov tree models," in *Proc. of IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2005, pp. 207–210.

[8] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] J. Berger, R. Coifman, and M. Goldberg, "Removing noise from music using local trigonometric bases and wavelet packets," *J. Audio Eng. Soc.*, vol. 42, no. 10, pp. 808–818, 1994.

[10] T. Painter, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–513, 2000.

[11] N. Vlassis and A. Likas, "A greedy EM algorithm for Gaussian mixture learning," *Neural Precessing Letters*, vol. 15, no. 1, pp. 77–87, 2002.

[12] S. Yoo, *Speech Decomposition and Speech Enhancement*, Ph.D. thesis, Department of Electrical and Computer Engineering, University of Pittsburgh, 2005.

[13] A. S. House, C. E. Williams, H. M. L. Hecker, and K. D. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Am.*, vol. 37, no. 1, pp. 158–166, 1965.

[14] C. Mackersie, A. C. Neuman, and H. Levitt, "A comparison of response time and word recognition measures using a word-monitoring and closed-set identification task," *Ear and Hearing*, vol. 20, no. 2, pp. 140–148, 1999.