SPEECH ENHANCEMENT BASED ON JOINT TIME-FREQUENCY SEGMENTATION

C. Tantibundhit[1,2]*, F. Pernkopf*[2]*, G. Kubin*[2]

[1]MedIntelligence and Innovation Laboratory, Thammasat University, Thailand
[2]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
tchartur@engr.tu.ac.th, pernkopf@tugraz.at, g.kubin@ieee.org

ABSTRACT

We present an algorithm to decompose speech into transient and non-transient components. Our algorithm, the joint time-frequency segmentation algorithm, uses the wavelet packet coefficients of the speech signal and represents them as tiles of a time-frequency representation adapted to the characteristics of the signal itself. Any wavelet packet coefficient, whose tiling height is larger than or equal to the tiling width is characterized as a transient coefficient and vice versa for the non-transient coefficient. The transient component is selectively amplified and recombined with the original speech to generate the modified speech with energy adjusted to be equal to the energy of the original speech. The psychoacoustic tests performed with fourteen human listeners show that the speech modification significantly improves speech intelligibility in background noise, i.e., for 10% absolute at 0dB to 31% absolute at -30dB.

Index Terms— Speech enhancement, transient component, speech intelligibility, wavelet packet transform

1. INTRODUCTION

During the past decades, there has been a vast increase in research focused on improving the intelligibility of speech presented in background noise, which can be divided into two categories. Speech enhancement of the first category aims to increase the intelligibility of speech already corrupted by noise by minimizing its effect as much as possible, e.g., active noise cancelation and spectral subtraction [1]. These approaches have been applied to noisy speech arrived at the listener, where the properties of noise, e.g., its spectrum are assumed to be available [1]. Although these approaches show impressive improvements, they may not work well under the conditions, where the noise is not known [2]. Speech enhancement of the second category are based on clean speech assumed to be available for processing before played back to a listener located in a noisy environment [2, 3]. The approaches are focused on the amplification of speech features shown to be important to speech perception, i.e., the transient components, without requiring the knowledge of background noise characteristics [2, 3].

Yoo *et al.* [2] developed an approach, where the original speech is first high-pass filtered at 700 Hz. Three time-varying bandpass filters are applied to capture the three strongest formants of high-pass filtered speech referred to as the quasi-steady-state (QSS) component. The QSS component is subtracted from the high-pass filtered speech resulting in the transient component. The transient component is selectively amplified and recombined with the original speech to generate the modified speech with the energy adjusted to be equal to the energy of the original speech. The intelligibility of the modified speech in background noise is compared to that of the original speech. The modified speech significantly improves speech intelligibility at low signal-to-noise ratios (SNRs), i.e., up to 32% at -25dB. However, the resulting transient component appears to retain a significant amount of formant energy during what would appear to be QSS regions of the speech and cannot capture the transient component frequencies below 700 Hz [3].

The approach of Tantibundhit *et al.* [3] decomposes speech into three components, i.e., tonal, transient, and residual components, respectively. The modified discrete cosine transform (MDCT) is used to capture constant or slowly varying frequency information in speech referred to as the tonal component. The wavelet transform is used to capture abrupt changes in speech referred to as the transient component. The residual component is expected to have small energy with a flat spectrum. The transient component is used to enhance speech intelligibility in background noise as done in [2]. The psychoacoustic test results have shown that the transient component significantly improves speech perception in background noise at low SNR levels (up to 18% at -25dB). Although, this approach decomposes the transient component more effectively than [2], i.e., removing vowel formants more effectively and emphasizing abrupt changes in time-frequency, the transient component suffered from pre-echo distortion artifacts of the MDCT [4] in tonal estimation. This may explain the lower improvements of speech intelligibility compared with the improvements by Yoo *et al.* [2].

Therefore, in this paper, we develop another approach to capture the transient component in speech signals more effectively. Specifically, first, we decompose the transient component directly from the original speech as in [3]. To

avoid pre-echo distortion artifacts of the MDCT in [3], we use the wavelet packet transform. Second, we use a multiresolution algorithm [5], where both time and frequency tilings are adapted directly to the characteristics of the speech signal itself instead of using fixed time-frequency tilings of the MDCT or the wavelet transform as in [3]. We believe that the multiresolution approach provides a more effective decomposition of the transient component of the speech signal. Our previous algorithm [5] allows to decompose the speech signal into two different components, i.e., the transient and non-transient components, respectively. The wavelet packet coefficients of the speech signal are represented as tiles of the time-frequency representation adapted both in time and frequency. The transient component is obtained using all of the wavelet packet coefficients, whose tiling heights are greater than or equal to the tiling widths, and vice versa for the non-transient component. In the following, details of the algorithm, examples of speech decomposition results, and the new method for the generation of modified speech are described in Section 2. The experimental setup (a modified rhyme test) used to evaluate the intelligibility of the modified speech and the original speech is described in Section 3. The test results are presented in Section 4. Implications of the results and future work are discussed in Section 5.

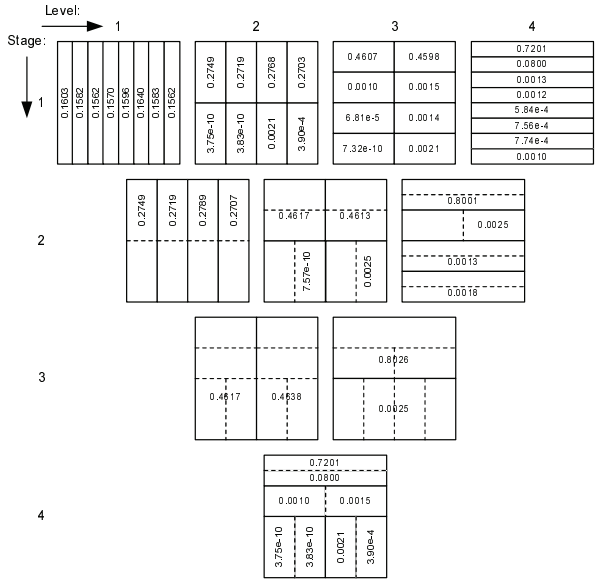

Fig. 1. Graphical representation of joint time-frequency segmentation for a 2048-sample synthetic signal composed of a high frequency (5 kHz) sinusoid and a single impulse.

2. SPEECH DECOMPOSITION AND MODIFICATION

2.1. Time-Frequency Representation

The original signal, $x_{\text{orig}}(t)$, sampled at 11.025 kHz, is transformed using the wavelet packet transform [6] limited to the coarsest level L composed of 256 coefficients (23.2 msec). The Daubechies-16 (Db16) wavelet is chosen as a mother wavelet because it gives a better estimation of the transient component across 300 monosyllabic consonant-vowel-consonant (CVC) rhyming words [7].

From the finest level (level 1) to the coarsest level (level L), the wavelet packet coefficients in each bin are divided into blocks of coefficients, each of which is composed of 256 coefficients. Then, all of the blocks of coefficients are windowed by the Hanning window based on the idea of Learned [8]. In the classification process, the use of all wavelet packet coefficients in the bin may lead to miss strong time-dependent features such as the transient information. Hence, it may be beneficial to calculate a windowed energy [8]. The window size of 128 coefficients (11.6 msec) with 50% overlap is chosen resulting in a half-window at the beginning and at the end of the block and three full windows, respectively. The average energy of each block of windowed coefficients is calculated resulting in five average energies in each block. Finally, the entropy of each block is calculated based on these average energies and is referred to as a cost of the coefficient block.

The next step is to evaluate all of the possible combinations of time-frequency tilings in every level (level 1 to level L) and find the combinations of time-frequency tilings that achieve the minimum cost. This can be achieved by performing the modified forward and modified backward algorithms explained in our previous work [5]. Figure 1 graphically shows how our algorithm works for a 2,048-sample synthetic signal composed of a high frequency (5 kHz) sinusoid and a single impulse located at the 1,345 sample. The tilings are expected to be split in frequency, where the 5 kHz sinusoid is located and expected to be split in time, where the single impulse is located. The number in each block of Fig. 1 represents the cost of this block of coefficients.

Starting from the first stage, and moving from level 1 to level L, the sum of the cost of two adjacent blocks in a considered level and the sum of the corresponding two transformed blocks (low-frequency and high-frequency) at the coarser level are compared. If the sum of the cost of two blocks in a considered level is less than or equal to the sum of the cost of the corresponding two transformed blocks, a time split is performed; otherwise, a frequency split is performed. The resulting time-frequency splits and the winning costs are put in the second stage. At this stage, the number of levels is reduced by one to $L-1$. The same approach is applied in the next stage with the number of levels reduced by one from the previous stage until reaching the last stage (stage L). At this stage, there is only one level left resulting in time-frequency tilings with a minimal cost, where the tilings are adapted both in time and in frequency to the characteristics of the analyzed signal itself.

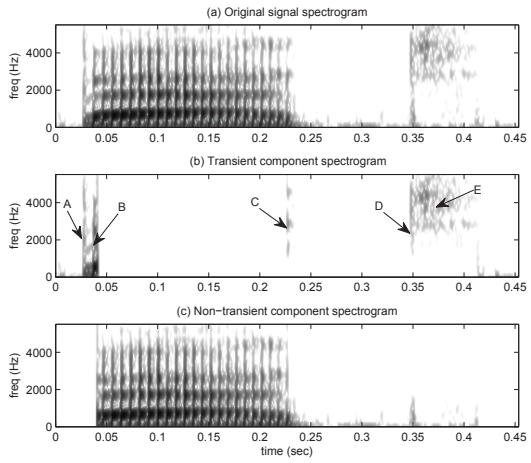

Fig. 2. Speech decomposition of "bat".

2.2. Transient Estimation

After the optimal time-frequency tiling has been achieved, the next step is to derive the transient component from the resulting tiling. All of the blocks of coefficients whose tiling height is more than the tiling width are characterized as transient tiles and vice versa for the non-transient tiles. All of the wavelet packet coefficients in the transient tiles, referred to as the significant wavelet packet coefficients, are retained but those in the non-transient tiles, referred to as the non-significant wavelet packet coefficients, are set to zero based on the idea of transform coding [3]. Then, the transient component, $x_{tran}(t)$, is estimated by the inverse wavelet packet transform of those significant wavelet packet coefficients. The non-transient component is calculated by subtraction of the transient component from the original speech signal as $x_{nont}(t) = x_{orig}(t) - x_{tran}(t)$.

2.3. Speech Decomposition Results

Figure 2 illustrates speech decomposition results for the mono-syllablic CVC word "bat", spoken by a male speaker. This word, phonetically transcribed as /bæt/, represents relatively simple distinctions between transient and non-transient components. Specifically, consonants, transitions from consonant to vowel, and transitions at the end of vowel are expected to be included in the transient component. Constant formant frequency information in vowels is expected to be included in the non-transient component. The spectrogram of the word is illustrated in Fig. 2a.

The transient component, illustrated in Fig. 2b, includes 3.8 % of the energy of the speech signal. It includes the release of the plosive /b/ (arrow A), the transition from /b/ to vowel /æ/ (arrow B) and the transition at the end of vowel /æ/ (arrow C), and most of the release of the plosive /t/ (arrow D). It also includes the aspiration noise of /t/ (arrow E) visible as a noise pattern in high frequency regions. The remaining non-transient component, illustrated in the bottom of the figure, includes most of the energy (96.2 %) of the speech signal. It predominantly includes the vowel /æ/ as expected.

2.4. Modified Speech to Improve Speech Intelligibility

The transient component is used to improve speech intelligibility, i.e., the transient component is selectively amplified and recombined with the original speech, with the total energy adjusted to be equal to the energy of the original signal based on the idea of [2, 3]. The transient amplification factor of 12 is chosen based on informal listening tests, which is the same factor as used in [2, 3]. A too small amplification factor results only in a small improvement of speech intelligibility while a large value results in a too strong emphasis of consonants and transitions in speech, leading to unnatural sounding speech and an implicit attenuation of the vowel sounds.

3. EXPERIMENTAL SET UP: MODIFIED RHYME TEST PROTOCOL

The objective of this experiment is to investigate whether the amplification of the transient component can improve the intelligibility of speech in background noise. This test protocol is a modified version of the word monitoring task of Mackersie *et al.* [9] using 300 monosyllabic CVC rhyming words proposed by House *et al.* [7].

The test protocol was performed on fourteen volunteer subjects with negative otologic histories and having at least one ear of hearing sensitivity of 15dB hearing level (HL) or better by conventional audiometry (250–8 kHz). Fifty sets of rhyming monosyllabic CVC words (6 words per set) were recorded by a male speaker as used in [2, 3]. Among them, 25 sets differ in their initial consonants and 25 sets differ in their final consonants. In each trial, subjects heard up to six acoustic stimuli corrupted by one level of speech-weighted background noise chosen randomly from six signal-to-noise ratio (SNR) levels (0, -6, -12, -18, -24, and -30dB). The target word appears as text on the computer display and remains visible until termination of the trial.

Subjects have to identify which stimulus is the target word. Subjects hear each stimulus only once and have to press the "SUBMIT" button as soon as they have recognized a stimulus as the target word. Then, the trial is terminated and the next trial is presented. If they think that the stimulus just heard is not the target word, they have to press the "NEXT" button to hear the next stimulus. The whole experiment is composed of one training session and three test sessions. Each test session is composed of one hundred trials. In this paper, we present first results of the experiment, i.e., the analysis of 75 trials of the original and 75 trials of the modified speech.

SNR	Mean difference	SD	95% CI
−30dB	31.18	19.26	20.06 to 42.30
−24dB	23.40	10.96	17.07 to 29.78
−18dB	26.69	9.69	21.10 to 32.29
−12dB	17.67	15.40	8.78 to 26.57
−6dB	0.14	13.64	−7.74 to 8.01
0dB	10.35	12.42	3.18 to 17.52

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

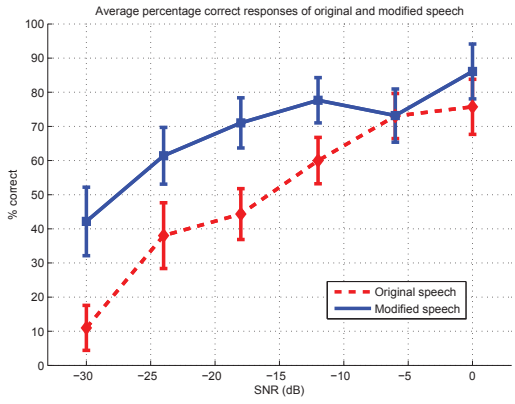

Fig. 3. Average percentage correct responses of original (dashed line) and modified speech (solid line).

4. PSYCHOACOUSTIC TEST RESULTS

The average percentage correct responses at each SNR level are calculated by the subjects' correct responses divided by the total number of stimuli. Means, standard deviations (SDs), and 95% confidence intervals (CIs) of the paired-sample difference at each SNR level are calculated and shown in Table 1. The results show that the modified speech is recognized better than original speech at all SNR levels with minimum improvement of 0.14% at −6dB and maximum improvement of 31.18% at −30dB. The modified speech significantly improves speech intelligibility in background noise in five of six SNR levels, i.e., 10% at 0dB, 18% at −12dB, 27% at −18dB, 23% at −24dB, and 31% at −30dB, respectively. At these SNR levels, the 95% CI differences do not include the value zero. However, the 95% CI difference at −6dB includes the value zero, an effect which still requires further study.

5. DISCUSSION

We have developed a joint time-frequency segmentation algorithm, where the tiling is adapted both in time and frequency based on the characteristics of the signal itself. The transient component is obtained using all of the wavelet packet coefficients, whose tiling heights are larger than the tiling widths. The transient component is used to enhance speech intelligibility in background noise.

The intelligiblity of the modified speech in background noise is better than that of the original speech for all six SNR levels suggesting that the transient component is important to speech perception. Our algorithm can improve speech intelligibility up to −30dB, while Yoo *et al.* [2] and Tantibundhit *et al.* [3] showed the improvements up to −25dB. Furthermore, our algorithm can improve speech intelligibility, even if the intelligibility is already high (above −10dB [3]). Specifically, speech intelligibility of the modified speech of Yoo *et al.* and Tantibundhit *et al.* is not better than that of the original speech at 0 and −5dB. Our modified speech significantly improves speech intelligibility in background noise at 0, −12, −18, −24, and −30dB, respectively. In future work, we will perform a direct experimental comparison of our new algorithm with the algorithms of Yoo *et al.* and Tantibundhit *et al.*.

6. REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton, FL, 2007.

[2] S.D. Yoo, J.R. Boston, A. El-Jaroudi, C.C. Li, J.D. Durrant, K. Kovacyk, and S. Shaiman, "Speech signal modification to increase intelligibility in noisy environments," *J. Acoust. Soc. Am.*, vol. 122, no. 2, pp. 1138–1149, 2007.

[3] C. Tantibundhit, J.R. Boston, C.C. Li, J.D. Durrant, S. Shaiman, K. Kovacyk, and A. El-Jaroudi, "New signal decomposition method based speech enhancement," *Signal Processing*, vol. 87, no. 11, pp. 2607–2628, 2007.

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

[5] C. Tantibundhit and G. Kubin, "Joint time-frequency segmentation for transient decomposition," in *Proc. of Interspeech*, Sep. 2008.

[6] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1998.

[7] 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.

[8] R.E. Learned, "Wavelet packet based transient signal classification," M.S. thesis, Department of Electrical Engineering, Massachusetts Institute of Technology, 1992.

[9] 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.