SPEECH ENHANCEMENT BASED ON JOINT TIME-FREQUENCY SEGMENTATION

C. T antibundhit1,2, F. Pernkopf2, G. Kubin2

1MedIntelligence and Innovation Laboratory, Thammasat University, Thailand
2Signal 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 tilling height is larger than or equal to the tilling 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.

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.

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

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

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 intelligibility 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