High Resolution Speech Feature Parametrization for Monophone-Based Stressed Speech Recognition
Ruhi Sarikaya and John H. L. Hansen

Abstract—This letter investigates the impact of stress on monophone speech recognition accuracy and proposes a new set of acoustic parameters based on high resolution wavelet analysis. The two parameter schemes are entitled wavelet packet parameters (WPP) and subband-based cepstral parameters (SBC). The performance of these features is compared to traditional Mel-frequency cepstral coefficients (MFCC) for stressed speech monophone recognition. The stressed speaking styles considered are neutral, angry, loud, and Lombard effect speech from the SUSAS database. An overall monophone recognition improvement of 20.4% and 17.2% is achieved for loud and angry stressed speech, with a corresponding increase in the neutral monophone rate of 9.9% over MFCC parameters.

Index Terms—Feature extraction, speech recognition, speech under stress, wavelet analysis.

I. INTRODUCTION

It is known that stress-based variations in speech production degrade the performance of speech recognition algorithms [9]. Stress refers to speech produced under environmental, emotional, or workload task. If the presence and type of stress could be determined, this extra information could be incorporated into recognition, coding, or synthesis algorithms to improve system performance [9]. Analysis, compensation, and utilization of stress in previous studies have generally focused on word level processing of speech and are not as easily generalized to continuous speech recognition. Improvement in large vocabulary recognition can be considered by investigating the effects of stress on phone recognition and proposing solutions for performance improvement. Previous studies have also considered wavelet-based analysis schemes such as subband-based cepstral (SBC) and wavelet packet parameters (WPP), which have been shown to outperform traditional features for stressed speech classification [1] and speaker identification [2].

The effects of monophone recognition accuracy across four types of stress have been previously investigated and shown to differ significantly between high energy voiced and low-energy consonant sections [8]. In this letter, the primary aim is to demonstrate significant gains in recognition rates across all stress conditions at no extra computational cost by using features obtained via wavelet analysis.

II. BASELINE MONOPHONE RECOGNITION SYSTEM AND DATABASE

The speech data employed in this study is from the speech under simulated and actual stress (SUSAS) database [7]. One part of SUSAS consists of talking style data such as neutral, angry, soft, loud, slow, fast, clear, and Lombard effect speech. Our data for each phone is extracted from the 35-word vocabulary using a previously formulated speech segmentation tool. Since the phonemes contained in the 35-word vocabulary do not span the entire 48 Texas Instruments/Massachusetts Institute of Technology (TIMIT) phone set, our phone list is limited to 29 phones. However, there is at least one phone from each of the 12 broad phone classes, as shown in Table I.

In this study, we consider only male speakers with all recognition evaluations performed in a speaker-dependent mode. A three-state, left-to-right, HMM-based recognizer is formulated with any mixtures per state. Phoneme boundary information is provided to the recognizer and hence, the experiments are isolated phoneme recognition. There are 12 available training tokens for each phone per speaker. Training and testing are conducted in a round-robin fashion, where each phone is trained with eleven tokens, leaving one out for testing. Overall performance is obtained as the average of ensemble results for all speakers. In the evaluation phase, there are a total of 3132 test tokens for the neutral monophones and 522 tokens for each of the three stress conditions. The ensemble of results over all nine speakers are averaged to obtain overall recognition rates.

III. TRADITIONAL AND WAVELET ANALYSIS-BASED FEATURES

The main difference between Mel-frequency cepstral coefficients (MFCC) and wavelet-based parameters lies in the computation of the spectrum. For MFCC’s, a discrete Fourier transform (DFT) is computed, whereas for the newly proposed parameters, a wavelet packet transform (WPT) is applied. The DFT samples the Fourier transform (FT) at equally spaced points in the frequency domain. In Fig. 1, a comparison of time-frequency tiling is shown for both DFT and WPT. WPT is computed via a time domain filtering with a subsignal representation obtained from frequency components within
TABLE I

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Neutral</th>
<th>Angry</th>
<th>Loud</th>
<th>Lombard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleus</td>
<td>MFCC</td>
<td>WPT</td>
<td>SBC</td>
<td>MFCC</td>
</tr>
<tr>
<td>Unvoiced Pratives</td>
<td>0.950</td>
<td>0.990</td>
<td>98.4</td>
<td>0.770</td>
</tr>
<tr>
<td>Voiced Pratives</td>
<td>0.720</td>
<td>0.970</td>
<td>95.9</td>
<td>0.550</td>
</tr>
<tr>
<td>Affricates</td>
<td>0.880</td>
<td>0.930</td>
<td>95.4</td>
<td>0.880</td>
</tr>
<tr>
<td>Vowels</td>
<td>0.850</td>
<td>0.970</td>
<td>95.9</td>
<td>0.850</td>
</tr>
<tr>
<td>Stops</td>
<td>0.070</td>
<td>0.110</td>
<td>97.2</td>
<td>0.070</td>
</tr>
<tr>
<td>Whisper</td>
<td>0.970</td>
<td>0.900</td>
<td>97.2</td>
<td>0.850</td>
</tr>
</tbody>
</table>

A. MFCC’s

MFCC’s [3] are the most common parametrization schemes employed in current recognition tasks. MFCC parameters are obtained in the frequency domain with critical band windowing of the speech spectrum.

B. WPP’s

A thorough discussion of wavelet analysis is beyond the scope of this paper, so we here refer readers to a more complete discussion in [5]. In continuous time, the wavelet transform is defined as the inner product of a signal $x(t)$ with a collection of wavelet functions $\psi_{ab}(t)$ in which the wavelet functions are scaled (by $a$) and translated (by $b$) versions of the prototype wavelet $\psi(t)$.

$$\psi_{ab}(t) = \psi \left( \frac{t-b}{a} \right)$$

$$W_\psi(x,a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^* \left( \frac{t-b}{a} \right) dt$$

Implementation of the wavelet transform (or more generally, the WPT) in discrete time is based on the iteration of a two-channel filterbank subject to certain constraints [4]. Unlike the wavelet transform, which is obtained by iterating on the low-pass filter branch, the filterbank tree can be iterated on either branch at any level, resulting in a tree-structured filterbank called a wavelet packet tree. Derivation of subband energies by this method rather than an FFT was first proposed in [6]. In this study, we employ a different wavelet packet tree that is a close approximation of the Mel-frequency division using Daubechies’s 32-tap orthogonal filters [2]. This structure better exploits the properties of the human auditory system. The subband signal energies are computed for each frame as

$$S_i = \frac{\sum_{m=1}^{M} \left| W_\psi(x,a,b) \right|^2}{N_i}$$

where $W_\psi(x,a,b)$ is the WPT of signal $x$, $i$ is subband frequency index ($i = 1, 2, \ldots, L$), and $N_i$ is the number of coefficients in the $i$th subband. The processing steps up to this point are the same.
for SBC and WPP. WPP’s are derived by taking the wavelet transform of the log-subband energies, where the parameters $a$ and $b$ in (2) are continuous. For discrete implementation, one often samples $\psi_{ab}(t)$ using $a = a_0^n$ and $b = n b_0 a_0^n$

$$\psi_{mn}(t) = a_0^{-m/2} \psi(a_0^{-m} t - nb_0) \quad (4)$$

where $m, n$ are integers. Therefore, WPP’s are the wavelet coefficients of the subband energies obtained as [2]

$$\text{WPP}(m, n) = \frac{1}{a_0} m^{2} \int \log S_t \psi(a_0^{-m} t - nb_0) \, dt. \quad (5)$$

It has been shown [4], [5] that multirate filters arranged in a dyadic tree can be used to compute the wavelet coefficients. In our case, we used Daubechies’s four tap filters to compute a three-level wavelet transform.

C. SBC’s

SBC parameters are derived by taking the discrete cosine transform of log-subband energies [1]. In this respect they are similar to MFCC, but the underlying method used to decompose the signal into subbands is different. We point out that in [6], an alternate set of SBC parameters are proposed from that in [1]. SBC parameters in [6] employ a different filterbank set and wavelet packet tree in combination with both static and delta parameters. Our SBC parameters are solely static features with filterbank and wavelet packet tree summarized in [1] and written as

$$\text{SBC}(n) = \sum_{i=1}^{L} \log S_t \cos \left( \frac{n(i-0.5)}{L} \pi \right), \quad n = 1, \ldots, N \quad (6)$$

where $N$ is the number of SBC parameters and $L$ is the total number of frequency bands. It was shown in [2] that WPP provides the highest speaker identification performance (followed by SBC) among these three feature sets on TIMIT data.

IV. RESULTS OF EXPERIMENTS

Monophone results from evaluations using SUSAS will be summarized within (intrafeature), as well as across feature domains (interfeature) for all four stress conditions (neutral, angry, loud, and Lombard) and the three parameter sets (MFCC, WPP, SBC) (see Table I).

A. MFCC

These results were previously reported in an earlier study [8] for monophone recognition of speech under stress. We summarize the main points here.

1) Neutral: Diphthongs, vowels, and semivowels that are usually long in duration resulted in very high recognition rates (93.3%). Consonant rates were not as easily recognized (75.8%) as high-energy voiced phone classes (HE-voiced). Rates were particularly low for stops.

2) Angry: If compared with neutral, angry recognition rates fell by $-44.7\%$ for consonants and $-60.8\%$ for HE-voiced.

Therefore, while the impact of the angry stress condition is high for all speech, it is more severe for HE-voiced.

3) Loud: If compared with neutral, recognition rates for loud fell by $-33.3\%$ and $-36.2\%$ for consonants and HE-voiced, respectively. Thus, the impact of loud condition is virtually equal for both broad phone classes but not as severe as that anger seen.

4) Lombard: Unlike angry and loud, the Lombard condition affected recognition performance of consonants ($15.8\%$) more than for HE-voiced ($11.6\%$). The overall impact was not as severe as that for angry and loud.

B. WPP

1) Neutral: HE-voiced achieved very high recognition rates (97.3%). Although consonant rates were not as high (91.4%), the difference is only 6.0\%, which shows that WPP improved consonant rates significantly over MFCC. Again, stops yielded the lowest individual consonant rate.

2) Angry: With respect to neutral, recognition rates fell by $-37.4\%$ for consonants and $-62.1\%$ for HE-voiced. Angry stress has a more pronounced impact on HE-voiced than consonants.

3) Loud: Performance resembles that experienced for angry, since both have a more severe impact on HE-voiced compared to consonants ($-36.6\%$ versus $-21.1\%$).

4) Lombard: Performance impact is similar to loud, as both resulted in virtually the same recognition reduction in broad phone classes.

C. SBC

1) Neutral: Again, HE-voiced attained higher recognition rates (97.2%) compared to consonants (89.4%). In particular, nasals, voiced fricatives, and whispers achieved very high rates.

2) Angry: Resulted in a $-35.4\%$ performance reduction for consonants and $-54.0\%$ for HE-voiced.

3) Loud: Rates fell by $-16.1\%$ for consonants and $-31.4\%$ for HE-voiced. However, the loss was not as severe as that for angry.

4) Lombard: Recognition rates were similar to that for the loud style.

D. Interfeature Observations

An overall comparison of WPP and SBC versus MFCC results is summarized at the bottom row of Table I. For consonants, SBC and WPP outperformed MFCC consistently for all stress styles including neutral. In the presence of stress, the high frequency content of the speech is often highly emphasized because of the change in spectral slope. This effect introduces more variability in high frequency bands, which represent much of the information for consonants. We attribute this to reduced local spectral variability in the wavelet representation because of stress. A better frequency localization is achieved compared to the short time FT for each of the subbands by using filters that have maximum vanishing moments.

Although SBC and WPP outperform MFCC for neutral, angry, and loud speech, MFCC’s perform better for Lombard speech in recognition of HE-voiced monophones.
It is encouraging that the new features achieved a 9–10% improvement over MFCC’s for neutral speech, which demonstrates their effectiveness for general speech recognition.

V. DISCUSSION AND CONCLUSIONS

It has been shown previously that stress has a nonuniform impact on speech production [8]. While front-end stress compensation is useful, the complexity is high [9]. The results here clearly show that high resolution-based features can measurably improve monophone-based speech recognition under stress with low processing complexity. Additional evaluations were performed using MFCC’s to explore the impact of stress on phones in different word position (i.e., initial, mid, final). The results showed that phone position did impact monophone rates [10] for the same stress conditions considered in Table I. In conclusion, we have explored the impact of stress on monophone recognition for neutral, angry, loud, and Lombard effect speech. Two alternative speech feature representations were considered based on wavelet analysis. It was shown that WPP and SBC parameters clearly outperform traditional MFCC’s for monophone-stressed speech recognition and that the performance improvement was more significant and consistent for consonant phone classes. It is believed that this is due to the nonuniform time-frequency resolution in wavelet-based methods. The results therefore suggest a viable parameter representation to improve stress robustness for large continuous vocabulary recognition.

REFERENCES