AN ALTERNATE FRONT-END BASED ON DIFFERENTIAL POWER SPECTRUM FOR SPEECH RECOGNITION

Kuldip K. Paliwal, Jingdong Chen and Xuechuan Wang

***Abstract*—The mel-scale frequency cepstral coefficients (MFCCs) derived from Fourier analysis and filter bank are perhaps the most widely used front-ends in state-of-the-art speech recognition systems. One of the major issues with the MFCCs is that they are very sensitive to additive noise. Hence many sorts of adaptation methods are often used to improve the robustness of a speech recognition system to the additive noise. In this paper, however, we propose a new feature set based on the differential power spectra and use them as an alternate front-end or a supplement to the MFCCs. We show that the proposed features are able to remove the additive noise in the spectral domain. The recognition results show that these features can improve the performances in both cases in which they are used either as independent features or as supplement features to the MFCCs.**

I. INTRODUCTION

Speech signal carries information from many sources. But not all information is relevant or important for a concrete task. In speech recognition, the first step often called feature analysis or front-end processing is designed to convert the speech signal into some acoustic features which hopefully only encapsulate the important information that is necessary for recognition. Once these features are computed, a post-end classifier is used to classify the input speech signal into certain pattern of class or sequence of classes in light of the extracted feature vectors and pre-trained models.

Different selection of acoustic features may greatly affect the performance of a speech recognizer. Feature selection should follow such a rule that the represented parameters should contain maximum information necessary for post-end classification and meanwhile, discard irrelevant information such as speaker variability, noise and channel distortions, etc. as much as possible so that the speech recognizer based on these parameters is robust to speaker variation and environment changes.

A great deal of work has been done for feature selection [1]. The mel-scale frequency cepstral coefficients (MFCCs) derived though Fourier analysis and from filter bank are perhaps the most commonly used acoustic features in currently available speech recognition systems. Much evidence has shown that the MFCCs have served as very successful front-ends for hidden Markov model (HMM) based speech recognition in the past decade [2]. Many speech recognition systems based on these front-ends have achieved very high level of accuracy in clean speech environment.

However, the MFCCs are also found to be sensitive to the noise distortion, especially the additive noise. To improve the robustness of a speech recognition system to additive noise, in the literature, various adaptation methods have been proposed such as spectral subtraction [8], Lin-log RASTA [3], parallel model composition (PMC) [4], etc. These adaptation methods take advantage of the prior knowledge of noise to mask, cancel or remove the noise during front-end processing or adjust the system parameters to match the new noisy environment and hence improve the system performance. There are also some papers focusing on extracting noise resistant feature set to improve the performance of a system [5].

In this paper, we want to investigate to use differential power spectrum as acoustic features for speech recognition. We show that the additive noise distortion can be removed by a differential operation in spectral domain if the noise is stationary or changing slower than speech signal. Just like the power spectrum can be converted to the MFCCs, we pass the differential power spectrum to a set of mel-scale frequency filter bank and convert them to some MFCC-like coefficients. The new features are shown to be more robust to the additive noise.

The proposed features can be used either as the independent features or a supplement to the MFCCs. Experiment results show that if the new features themselves are served as front-ends, their performance is better than the MFCCs; if the new features are used as a supplement to the MFCCs, they can improve the performance of a speaker-independent speech recognizer.

II. CEPSTRUM-LIKE FEATRES BASED ON DIFFERENTIAL POWER SPECTRUM

If $s(t)$ is the original clean speech signal, the received speech signal $y(t)$ is modeled as

$$y(t) = s(t) * h(t) + n(t) \tag{1}$$

where $h(t)$ is the impulse response of channel distortion and $n(t)$ the ambient noise. $*$ denotes the convolution operation.

Speech signal is time-variant and non-stationary, it is usually analyzed on the frame-by-frame basis. For one frame of speech signal, the equation (1) is written as

$$y(k,t) = s(k,t) * h(k,t) + n(k,t) \tag{2}$$

Were index k denotes the k^{th} frame. Assume that the noise in (2) is uncorrelated with speech signal, the power spectrum of the above received speech signal is

$$P_y(k,f) = P_s(k,f)\left|H(k,f)\right|^2 + P_n(k,f) \quad (3)$$

Compared with speech signal, the channel distortion and effect of noise vary much slower. Hence the power spectrum of the k^{th} frame of speech signal can be rewritten as

$$P_y(k,f) = P_s(k,f)\left|H(f)\right|^2 + P_n(f) \quad (4)$$

In speech recognition, the power spectrum is often converted to the MFCCs as front-ends. Passing the power spectrum of a speech signal as shown in equation (4) to a set of triangle filters which are equaly spaced in mel-scale freqency axis gives the filterbank engies. These can be tranformed to MFCCs by applying a discrete cosine transform (DCT) to the logarithm of the fitler bank engies.

The MFCCs are widely used as front-end parameters in most current speech recognition systems. They are proven to have very high discriminities in clean speech case. Additionally convolutionary distortions which appear as complicative componets to speech in the power spectrum are shown to be additive components to the speech parameters in cepstral domain and hence much easier to be removed in cepstral domain.

Additive noise, which is an additive part in spectral domain, however, behaves as multiplicative distortions in cepstral domain [6] which is not easy to be removed and hence cause the MFCCs be sensitive to additive distortions.

In this paper, We introduce differential power spectrum (DPS), which is derived from (4) and defined as

$$P_y^D(k,f) = P_y(k+1,f) - P_y(k,f) \quad (5)$$

Substituting the power spectrum as defined in (4) to (5) we get

$$P_y^D(k,f) = \left|H(f)\right|^2 [P_s(k+1,f) - P_s(k,f)] \quad (6)$$

The above equation indicates that the additive noise distortion, which appears as an additive part in power spectrum domain, is removed in DPS and hence the DPS should be more robust to additive noise.

Just like the power spectrum can be converted to the MFCCs, the DPS can be transformed to some MFCC-like coefficients. Passing the DPS to a mel-scale frequency triangle filter bank gives a set of energy-like outputs. We do not call the outputs the fitler bank energies since the DPS does not directly reflect the energy distribution of a singal in frequency domain. Another reason we call them energy-like outputs because energy has non-negative value. However, because the DPS may have negative values, the filter bank outputs often have negative values.

A log-operation is also necessary for compressing the dynamic range of the energy-like filter bank outputs. However, Since the filter bank outputs have negative values, the log-operation used in the estimation of the MFCCs is noy usable in such case. we redefine a log-operation as follows. For the k^{th} frame of speech signal, passing its differential power spectrum to the fitler bank gives the outputs which are denoted as $E[k,n]$, where $n = 1,2,\cdots,N$, N is the total number of filters. The log-operation to the outputs is defined as

$$\log E[k,n] = \log\left|E[k,n]\right| + i\arg E[k,n] \quad (7)$$

where

$$\arg E[k,n] = \begin{cases} 0, & if\ E[k,n] \geq 0 \\ \pi, & if\ E[k,n] < 0 \end{cases} \quad (8)$$

After the log-operation, a DCT may applied to the logrithm energy-like outputs to convert them to some MFCC-like coefficients. The DCT may be applied to the complex data, real parts of the data or imaginary parts of the data. These may result different configurations of the feature extractors which are shown in the following section.

III. RECOGNITION RESULTS

Experiments have been done to evaluate the performances of the proposed features. The recognition system used is an HMM-based speaker-independent isolated speech recognizer. Models are left-to-right with no skip state transition. Eight states are used for each model and training iterations begin with uniformly probabilistic model. The vocabulary consists of 26 English alphabets.

The database used is TI46 which contains 16 speakers including 8 males and 8 females. There are 26 utterances of each word from each speaker. 10 of them are designated as training tokens and the rest 16 are designated as testing tokens. Speech is digitized at a sampling rate of 12.5kHz with 12-bit quantization value for each sample.

Speech signals in this database are corrupted by very strong noise. One exemplary speech waveform is shown in Fig.1. The average signal-to-noise ratio (SNR) of the whole alphabet database is about 3dB. (The definition of the SNR we used is defined as $SNR = 10\log_{10}(\frac{\sigma_{s+n}^2 - \sigma_n^2}{\sigma_n^2})$).

In the front-end processing, the speech signal is analyzed every 15ms with a frame width of 30 ms (with Hamming window and preemphasis). The mel-scale frequency filter bank consists of 25 triangle filters.

We have twelve different configurations of features shown as follows:

Configuration 1: 12 MFCCs are used as front-ends.

Configuration 2: The logarithm energy-like filter bank outputs are complex data that contain real parts and imaginary parts. In this configuration, DCT is applied to only real parts of the outputs and convert them to 12 coefficients.

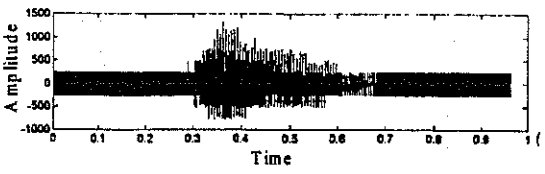

Fig. 1 Speech waveform for 'A' (speaker is m1)

Configuration 3: The logarithm energy-like filter bank outputs are complex data. In this configuration, A DCT is applied to the moduli of these complex data and converted them to 12 coefficients.

Configuration 4: The real parts and imaginary parts of logarithm energy-like outputs are converted to 12 coefficients separately by applying a DCT. 24 coefficients are used as front-ends in this configuration.

Configuration 5: This Configuration is the combination of Configuration 1 and Configuration 2. 24 coefficients are used in this case.

Configuration 6: This Configuration is the combination of Configuration 1 and Configuration 3. 24 coefficients are used in this case.

Configuration 7: 12 MFCCs plus 12 delta MFCCs that are calculated by subtracting the two preceding from the two following vectors are used as features.

Configuration 8: 12 coefficients in Configuration 2 plus 12 delta coefficients that are derived from Configuration 2 with the use of the same procedure to calculate the delta MFCCs are used as front-ends.

Configuration 9: 12 coefficients in Configuration 3 plus 12 delta coefficients that are derived from Configuration 3 with the use of the same procedure to calculate the delta MFCCs are used as front-ends.

Configuration 10: 24 coefficients in Configuration 4 plus 24 delta coefficients that are derived from Configuration 4 are combined together to construct a feature vector of 48 dimension as front-ends. The delta features are calculated by subtracting two preceding from the two following static feature vectors.

Configuration 11: This configuration derives from Configuration 5, which includes 24 statistic features and 24 delta coefficients.

Configuration 12: This configuration derives from Configuration 6, which includes 24 statistic features and 24 delta coefficients.

The recognition results are shown in Table 1 and Table 2. The column 2 of these tables lists the number of parameters used in the feature representation. The column 3 lists the recognition rates for different configuration of features. Many observations can be seen from the results.

1. If the proposed features are used as alternate front-ends, they perform better that the MFCCs.

For male speakers, with the use of 12 static MFCCs, we get an accuracy of 76.9% (Conf. 1). However, if the differential power spectrum is used as the input of filter bank, only real parts of the outputs are used, the recognition rate is 78.5% (Conf. 2). If moduli are used, the recognition accuracy is 79.1% (Conf. 3). If real parts and imaginary parts are converted to MFCC-like coefficients separately and combine them together as the front-ends, the recognition rate is 81.9% (Conf. 4). Comparing with the MFCCs, the error reductions for these three cases are 6.9%, 9.5% and 21.6% respectively.

For gender-independent recognition task, the similar results are observed (Table 2, Conf. 2 ~ Conf. 4). The error reductions for the three cases are 1.6%, 11.0% and 39.1%.

2. If the proposed features are used as a supplement to the MFCCs, they can greatly improve the recognition performance.

From Table 1, it can be seen that only use the 12 MFCCs as front-ends, the recognition rate is 76.9% (Conf. 1). If 12 coefficients derived from the real parts of the logarithm energy-like filter bank outputs are used as a supplement, the recognition rate raises to 79.7% (Conf. 5). The corresponding error reduction is 12.1%. If 12 coefficients transformed from the modulus of the logarithm energy-like filter bank outputs are used a supplement, the recognition accuracy raises to 81.1% (Conf. 6). The error reduction reaches 18.2%. Similarly, for gender-independent recognition task, the use of 12 MFCCs gives an accuracy of 68.3%. While combining 12 MFCCs and 12 coefficients estimated from the real parts of the logarithm energy-like filter bank outputs together improves the recognition accuracy to 75.2%, the error reduction is 21.8%. If 12 MFCCs, together with 12 coefficients derived from the modulus of the logarithm energy-like filter bank outputs are used as front-ends, the recognition rate raises to 75.5%. The corresponding error reduction is 22.7%.

3. Dynamic MFCCs are often used to improve the recognition accuracy (Conf. 7). Similarly, the differentials of the proposed features are also useful for further improvement the performance, which can be seen from Conf. 8 ~ Conf. 10.

4. Combining the MFCCs, delta MFCCs, the proposed features and the differentials of the proposed features together yields the highest recognition accuracy in our task, which can be observed from Conf. 11, Conf. 12 in the Table 1 and Table 2.

5. If the proposed parameters are used as a supplement to the MFCCs, the dimensionality of the feature space is doubled. However, the linear discriminant analysis may be used for dimensionality reduction [7].

Configuration	Dimension of feature vector	Recognition rate (%)
Conf. 1	12	76.9
Conf. 2	12	78.5
Conf. 3	12	79.1
Conf. 4	24	81.9
Conf. 5	24	79.7
Conf. 6	24	81.1
Conf. 7	24	86.4
Conf. 8	24	83.0
Conf. 9	24	83.3
Conf. 10	48	85.9
Conf. 11	48	87.8
Conf. 12	48	87.0

Table 1. Recognition rates for different configurations of features. Recognizer is a continuous density HMM-based speaker-independent one but only male speakers are used for training and testing in this case. The PDF is represented by a mixture of 4 Gaussian densities.

Configuration	Dimension of feature vector	Recognition rate (%)
Conf. 1	12	68.3
Conf. 2	12	68.8
Conf. 3	12	71.8
Conf. 4	24	76.7
Conf. 5	24	75.2
Conf. 6	24	75.5
Conf. 7	24	79.56
Conf. 8	24	76.94
Conf. 9	24	78.2
Conf. 10	48	80.5
Conf. 11	48	79.6
Conf. 12	48	80.0

Table 2. Recognition rates for different configurations of features. Recognizer is a continuous density HMM-based gender-independent one in which all male and female speakers are used for training and testing. The PDF is represented by a mixture of 8 Gaussian densities.

IV. CONCLUSION

In this paper, differential power spectrum is investigated as front-end for HMM-based speech recognition. The differential power spectrum is converted to some MFCC-like acoustic features. Since the differential power spectrum is able to remove the additive noise in spectral domain, the proposed feature set is shown more robust to ambient noise. Recognition results based on a speaker-independent isolated speech recognizer show that performances improve greatly either when the proposed features are used as alternate front-ends or when they are used as a supplement to the MFCCs.

4. REFERENCES

[1] J. W. Picone, "Signal Modeling Technique in speech Recognition", Proc. IEEE, Vol. 81, No. 9, September 1993.

[2] Steve J. Young, "Large Vocabulary Continuous Speech Recognition: A review", IEEE Signal Processing, 1996.

[3] Hynek Hermansky and Nelson Morgan, "RASTA processing of Speech", IEEE Trans. On Speech and Audio Processing, Vol. 2, No. 4, Oct. 1994. PP. 578-589.

[4] M. J. F. Gales and S. J. Young, "Robust Speech Recognition in Additive and Convolutional Noise Using parallel Model Combination", Computer Speech and Language (1995)9, PP. 289-307.

[5] Kuldip K. paliwal, "Spectral Subband Centroid Features for Speech Recognition", ICASSP'98, PP. 617-620.

[6] Xiao Yu Zhang et al, "Channel and Noise Normalization Using Affine Transformed Cepstrum", ICSLP'96, Vol. IV, PP. 1993-1996, October 1996.

[7] Kuldip K. Paliwal, "Dimensionality Reduction for the Enhanced Feature Set for the HMM-based Speech Recognition", Digital Signal Processing, Vol. 2, No. 3, PP. 57-173, July 1992.

[8] A. Nolazco Flores and S. J. Young, "Continuous Speech Recognition in Noise using Spectral Subtraction and HMM Adaptation", ICASSP'94, Vol. I, PP.409-412, April 1994.