PRODUCT OF POWER SPECTRUM AND GROUP DELAY FUNCTION FOR SPEECH RECOGNITION

Donglai Zhu and Kuldip K. Paliwal

School of Microelectronic Engineering
Griffith University
Brisbane, QLD 4111, Australia
donglai.zhu@ustc.edu , k.paliwal@griffith.edu.au

ABSTRACT

Mel-frequency cepstral coefficients (MFCCs) are the most widely used features for speech recognition. These are derived from the power spectrum of the speech signal. Recently, the cepstral features derived from the modified group delay function (MGDF) have been studied by Murthy and Gadde [6] for speech recognition. In this paper, we propose to use the product of the power spectrum and the group delay function (GDF), and derive the MFCCs from the product spectrum. This spectrum combines the information from the magnitude spectrum as well as the phase spectrum. The MFCCs of the MGDF are also investigated in this paper. Results show that the cepstral features derived from the power spectrum perform better than that from the MGDF, and the product spectrum based features provide the best performance.

1. INTRODUCTION

Currently, the cepstral features are the most widely used features for speech recognition [1][2][3]. These features are derived from the power spectrum of the speech signal, while the phase spectrum is ignored. This is done mainly due to our traditional belief that the human auditory system is phase-deaf, i.e., it ignores phase spectrum and uses only magnitude spectrum for speech perception. Recently, it has been shown that the phase spectrum is useful in human speech perception [4]. This suggests that meaningful recognition features can be derived from the phase spectrum of the signal.

Some features derived from the phase spectrum have been studied in the literature [5][6]. In [5], instantaneous frequencies derived from the phase spectrum were proposed as features and were shown to give performance comparable with Mel-frequency cepstral coefficient (MFCC) features. In [6], Murthy and Gadde have modified the group delay function (GDF) to suppress the zeroes caused by pitch peaks, noise and window effects, and applied the discrete cosine transform (DCT) on the modified GDF (MGDF) to get the cepstral coefficients. We call them modified-group-delay cepstral coefficients (MGDCCs).

Since the Fourier transform of the speech signal is composed of the magnitude spectrum and the phase spectrum, the features derived from either the power spectrum or the phase spectrum have the limitation in representation of the signal. In this paper, we define the product spectrum as the product of the power spectrum and the GDF. It combines the magnitude spectrum and the phase spectrum. We derive the MFCCs from the product spectrum, and name them Mel-frequency product spectrum cepstral coefficients (MFPSCCs). We also investigate the MFCCs derived from the MGDF, named Mel-frequency modified-group-delay cepstral coefficients (MFMGDCCs). Results show that the MFMGDCCs and the MGDCCs are much worse than the MFCCs; the MFPSCCs give the best results.

2. PRODUCT SPECTRUM

Given a frame of speech signal $x(n), n = 0 \cdots N-1$, the Fourier transform is given by

$$X(\omega) = |X(\omega)| e^{j\theta(\omega)} \tag{1}$$

The GDF is defined as

$$\tau_p(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{2}$$

Equation (2) can be simplified as follows [7],

$$\begin{aligned} \tau_p(\omega) &= -\operatorname{Im}\frac{d(\log X(\omega))}{d\omega} \\ &= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \end{aligned} \tag{3}$$

where $Y(\omega)$ is the Fourier transforms of $nx(n)$, and the subscripts R and I denote the real and imaginary parts.

Figure 1 (a), (b) and (c) show a frame (of duration T =30ms) of the vowel sound /i/, its power spectrum and

GDF, respectively. Before the Fourier transform, the speech signal has been pre-emphasized and multiplied with Hamming window. In the power spectrum, the formants are clearly visible. However, there are only meaningless peaks and valleys in the GDF. It occurs due to the power spectrum in the denominator in Equation (3). In order to make the GDF meaningful, a modification to the GDF has been proposed by replacing the power spectrum $|X(\omega)|^2$ with the cepstrally smoothed power spectrum $(S(\omega))^2$ in Equation (3) [8]. This gives the MGDF as follows:

$$\tilde{\tau}_p(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{(S(\omega))^2} \qquad (4)$$

Figure 1 (d) shows the MGDF of the signal. Since the MGDF has negative values, it needs to be clipped by a nonnegative floor before the calculation of the dB values. We adopt the dynamic range threshold [2], i.e., discarding the values below a certain threshold from the peak in the spectrum. Here the threshold is set as $-60dB$. In the MGDF shown in the figure, the first formant is visible to some extent, but the other formants are lost. Also, the MGDF has a rather flat envelope, which is caused by the presence of the smoothed power spectrum term in the denominator in Equation (4).

In this paper, we define the product spectrum $Q(\omega)$ as the product of the power spectrum and the GDF as follows:

$$\begin{aligned} Q(\omega) &= |X(\omega)|^2 \tau_p(\omega) \\ &= X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \end{aligned} \qquad (5)$$

The product spectrum is influenced by both the magnitude spectrum and the phase spectrum. Figure 1 (e) shows the product spectrum of the signal. It enhances the region at the formants over the MGDF and has an envelope comparable to that of the power spectrum.

3. COMPUTATION OF RECOGNITION FEATURES

In order to investigate the performance of the product spectrum for speech recognition, we derive the MFCCs from the product spectrum, namely MFPSCCs. We compare the MFPSCCs with the following three features:

1. MFCCs derived from the product spectrum. We still call them MFCCs.
2. MGDCCs proposed in [6].
3. MFCCs derived from the MGDF. We call them MFMGDFCCs.

The following sections present the computation of these four features.

3.1. Mel-frequency cepstral coefficients

The MFCCs are computed in the following four steps [2]:

1. Compute the fast Fourier transform (FFT) spectrum of $x(n)$, denoted by $X(k)$.
2. Compute the power spectrum $|X(k)|^2$.
3. Apply a Mel-frequency filter-bank to $|X(k)|^2$ to get filter-bank energies (FBEs).
4. Compute DCT of log FBEs to get the MFCCs.

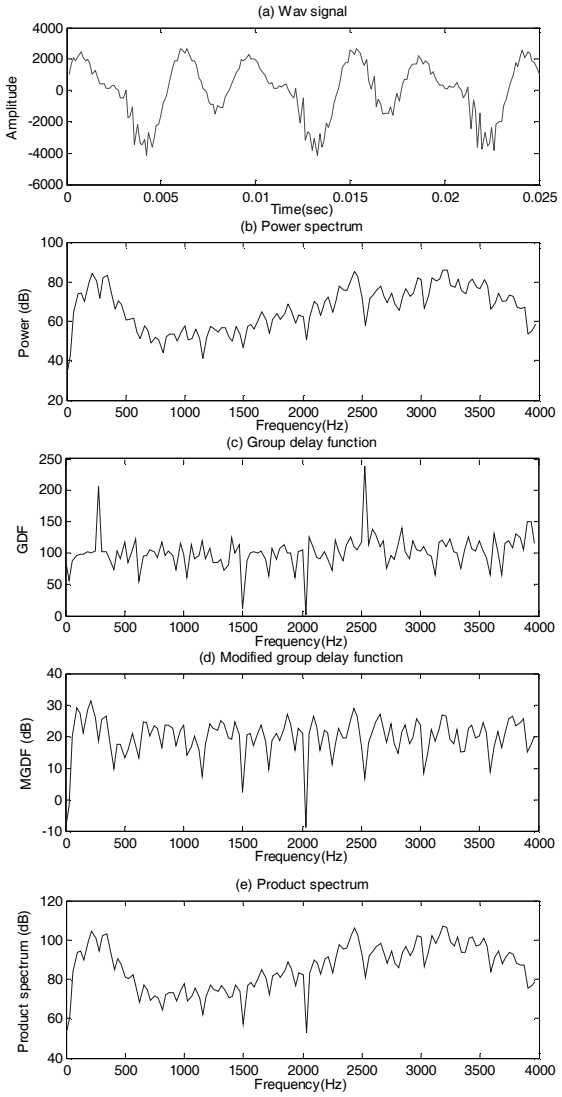

Fig 1: A frame of vowel sound /i/, its power spectrum, group delay function (GDF), modified group delay function (MGDF) and product spectrum

3.2. Modified-group-delay cepstral coefficients

The MGDCCs are computed in the following four steps[6]:

1. Compute the FFT spectrum of $x(n)$ and $nx(n)$. Denote them by $X(k)$ and $Y(k)$.
2. Compute the cepstrally smoothed spectrum of $|X(k)|$. Denote it by $S(k)$.
3. Compute the MGDF as follows:

$$\tilde{\tau}_p(k) = sign \cdot \left| \frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(k))^{2\gamma}} \right|^{\alpha} \quad (6)$$

where $sign$ is the sign of $\frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(k))^{2\gamma}}$.

4. Compute the DCT of $\tilde{\tau}_p(k)$ to get the MGDCCs.

3.3. Mel-frequency modified-group-delay cepstral coefficients

The MFMGDCCs are computed in the following five steps:

1. Compute the FFT spectrum of $x(n)$ and $nx(n)$. Denote them by $X(k)$ and $Y(k)$.
2. Compute the cepstrally smoothed spectrum of $|X(k)|$. Denote it by $S(k)$.
3. Compute the MGDF

$$\tilde{\tau}_p(k) = \max\left(\frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(k))^2}, \rho\right) \quad (7)$$

where

$$\rho = 10^{\frac{\sigma}{10}} \cdot \max\left(\frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(k))^2}\right) \quad (8)$$

σ is the threshold in dB.

4. Apply a Mel-frequency filter-bank to $\tilde{\tau}_p(k)$ to get the FBEs.
5. Compute DCT of log FBEs to get the MFMGDCCs.

3.4. Mel-frequency product-spectrum cepstral coefficients

The MFPSCCs are computed in the following four steps:

1. Compute the FFT spectrum of $x(n)$ and $nx(n)$. Denote them by $X(k)$ and $Y(k)$.
2. Compute the product spectrum

$$Q(k) = \max(X_R(k)Y_R(k) + X_I(k)Y_I(k), \rho) \quad (9)$$

where

$$\rho = 10^{\frac{\sigma}{10}} \cdot \max(X_R(k)Y_R(k) + X_I(k)Y_I(k)) \quad (10)$$

σ is the threshold in dB.

3. Apply a Mel-frequency filter-bank to $Q(k)$ to get the FBEs.
4. Compute DCT of log FBEs to get the MFPSCCs.

4. SPEECH RECOGNITION EXPERIMENTS

The Aurora2 database [8] was used to evaluate the performance. This database can be used to evaluate performance of speech recognition algorithms in noisy conditions. The source speech for this database is the TIDigits, consisting of connected digits task spoken by American English talkers, sampled at 8 kHz. There are two training sets (clean training set and multi-condition training set) and three test set. Test set A and B have speech corrupted by different real-world additive noises at the SNRs from -5 dB to 20dB at the step of 5dB. Test set C is influenced by both additive noise and convolutional noise.

In our experiments, we used the clean training set to train the HMMs, which were defined and trained in the same way as the Aurora2 baseline system [8]. In the calculation of all the features, the speech signal was analyzed every 10 ms with a frame width of 30 ms (with Hamming window and pre-emphasis). The Mel filter bank was designed with 23 frequency bands in the range from 64 Hz to 4 kHz. Finally, 12 cepstral coefficients were obtained. Cepstral mean subtraction was performed for all features. In the calculation of the MGDF, the cepstrally smoothed spectrum was derived from 13 lower-order cepstral coefficients (including the coefficient of order 0). In the calculation of the MGDCCs, the two parameters were set as $\alpha = 0.4$ and $\gamma = 0.9$. In the calculation of MFMGDCCs and the MFPSCCs, the threshold was defined as $\sigma = -60dB$.

Tables 1 and 2 show the accuracies of the features on the three test sets. In Table 1, the features only include 12 cepstral coefficients. In Table 2, the features include 12 cepstral coefficients, energy, delta and accelerator coefficients, totally 39 coefficients. For each test set, the accuracies are averaged over different noises. The last column is the average over the SNRs between 20dB and 0dB. From the results we may draw the following conclusions:

1. The MFCCs provide better performance than the MGDCCs and the MFMGDCCs. It indicates that the power spectrum gives better performance than the phase spectrum.
2. The MFMGDCCs obtain better performance than the MGDCCs at low SNRs, but worse at high SNRs or in clean condition. It indicates that the spectral smoothing with the Mel-frequency filter-bank is useful to derive a robust representation for mismatched conditions, but not helpful in matched condition.
3. The MFPSCCs obtain the best performance. It indicates that the product spectrum is better than the

power spectrum and the phase spectrum for speech recognition.

5. CONCLUSIONS

In this paper, we have introduced the product spectrum as the product of the power spectrum and the GDF. The product spectrum combines the information from the magnitude spectrum as well as the phase spectrum of the speech signal. We derived the MFCCs from the product spectrum (i.e., MFPSCCs), and compared them with the cepstral features from the phase spectrum (i.e., MGDCCs and MFMGDCCs) and the MFCCs from the power spectrum. Results showed that the power spectrum gives better performance than the phase spectrum; the product spectrum gives the best performance.

6. REFERENCES

[1] S. Young, "A Review of Large-Vocabulary Continuous Speech Recognition", *IEEE Signal Processing Magazine*, pp. 45-57, 1996

[2] J.W. Picone, "Signal Modeling Techniques in Speech Recognition", *Proc. IEEE*, vol. 81, No. 9, 1993

[3] L. Rabiner and B.H. Juang, *"Fundamentals of Speech Recognition"*, Prentice Hall, Englewood Cliffs, NJ, 1993

[4] K.K. Paliwal and L. Alsteris, "Usefulness of Phase Spectrum in Human Speech Perception", *Proc. Eurospeech*, pp. 2117-2120, 2003

[5] K.K. Paliwal and B.S. Atal, "Frequency-Related Representation of Speech", *Proc. Eurospeech*, pp. 65-68, 2003

[6] H.A. Murthy and V. Gadde, "The Modified Group Delay Function and Its Application to Phoneme Recognition", *Proc. ICASSP*, vol. 1, pp. 68-71, 2003

[7] A.V. Oppenheim and R.W. Schafer, "*Digital Signal Processing*", Englewood Cliffs, NJ: Prentice-Hall, 1975

[8] B. Yegnanarayana and H.A. Murthy, "Significance of Group Delay Functions in Spectrum Estimation", *IEEE Trans. Signal Processing*, vol. 40, pp. 2281-2289, 1992

[9] H.G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*; Paris, France, September 18-20, 2000

Test set	Feature set	SNR(dB)							
		Clean	20	15	10	5	0	-5	Ave
A	MFCC	96.64	88.71	79.00	60.14	34.88	18.69	12.86	56.28
	MGDCC	81.32	70.22	60.53	47.00	28.49	10.85	-0.64	43.42
	MFMGDCC	55.32	50.69	48.43	42.02	32.47	19.90	12.09	38.70
	MFPSCC	96.41	89.11	80.08	63.17	37.62	19.98	13.38	57.99
B	MFCC	96.64	90.14	82.98	66.87	41.94	22.54	13.75	60.90
	MGDCC	81.32	69.46	60.76	47.09	28.60	11.61	-0.16	43.50
	MFMGDCC	55.32	48.07	45.56	38.14	28.23	16.99	9.71	35.40
	MFPSCC	96.41	90.40	83.30	68.61	44.82	23.94	14.20	62.22
C	MFCC	96.65	89.82	80.67	62.65	38.80	20.97	14.11	58.58
	MGDCC	81.46	68.66	54.76	36.95	17.74	1.38	-5.16	35.90
	MFMGDCC	52.47	51.86	49.23	43.35	34.46	22.56	12.86	40.29
	MFPSCC	96.32	90.41	81.59	65.33	42.18	22.66	14.49	60.43

Table 1: Comparison of the features composed of only cepstral coefficients

Test set	Feature set	SNR(dB)							
		Clean	20	15	10	5	0	-5	Ave
A	MFCC+E+Δ+ΔΔ	99.34	97.04	92.24	76.79	44.70	22.36	13.04	66.63
	MGDCC+E+Δ+ΔΔ	86.50	75.27	65.60	50.71	27.69	4.17	-10.26	44.69
	MFMGDCC+E+Δ+ΔΔ	81.06	73.76	68.96	61.09	45.39	25.73	13.20	54.98
	MFPSCC+E+Δ+ΔΔ	99.31	96.94	92.36	78.68	48.60	23.37	13.43	67.99
B	MFCC+E+Δ+ΔΔ	99.34	97.81	94.28	82.73	54.48	26.93	14.26	71.25
	MGDCC+E+Δ+ΔΔ	86.50	73.58	64.37	48.84	26.07	5.06	-8.94	43.58
	MFMGDCC+E+Δ+ΔΔ	81.06	73.01	67.47	57.70	41.40	20.64	10.57	52.04
	MFPSCC+E+Δ+ΔΔ	99.31	97.77	94.22	84.08	57.86	28.45	14.66	72.48
C	MFCC+E+Δ+ΔΔ	99.27	96.65	90.82	74.05	43.44	21.96	13.59	65.38
	MGDCC+E+Δ+ΔΔ	86.81	71.78	56.00	34.59	8.34	-13.91	-26.44	31.36
	MFMGDCC+E+Δ+ΔΔ	80.65	72.55	67.76	59.11	43.73	24.17	12.20	53.46
	MFPSCC+E+Δ+ΔΔ	99.28	96.73	91.43	76.31	46.82	23.29	13.99	66.91

Table 2: Comparison of the features composed of cepstral coefficients, energy, delta and accelerator coefficients