THE MODIFIED GROUP DELAY FUNCTION AND ITS APPLICATION TO PHONEME RECOGNITION

Hema A Murthy

Department of
Computer Science & Engineering,
Indian Institute of Technology Madras,
Chennai - 600 036, India
e-mail:hema@TeNeT.res.in

Venkata Gadde

Speech Technology and Research Laboratory,
SRI International,
333, Ravenswood Avenue,
Menlo Park, CA 94025, USA
e-mail : rao@speech.sri.com

ABSTRACT

We explore a new spectral representation of speech signals through group delay functions. The group delay functions by themselves are noisy and difficult to interpret owing to zeroes that are close to the unit circle in the z-domain and these clutter the spectra. A new *modified group delay function* [1] that reduces the effects of zeroes close to the unit circle is used. Assuming that this new function is minimum phase, the modified group delay spectrum is converted to a sequence of cepstral coefficients. A preliminary phoneme recogniser is built using features derived from these cepstra. Results are compared with those obtained from features derived from the traditional mel frequency cepstral coefficients (MFCC). The baseline MFCC performance is 34.7%, while that of the best modified group delay cepstrum is 39.2%. The performance of the composite MFCC feature, which includes the derivatives and double derivatives, is 60.7%, while that of the composite modified group delay feature is 57.3%. When these two composite features are combined, $\approx$ 2% improvement in performance is achieved (62.8%). When this new system is combined with linear frequency cepstra (LFC) [2], the system performance results in another $\approx$ 0.8% improvement (63.6%).

1. INTRODUCTION

The objective of this study is to explore a set of new features based on the Fourier transform phase of a signal, rather than the conventional Fourier transform magnitude for speech recognition. The method described is based on the negative derivative of the Fourier transform (FT) phase function, also called the *group delay function*. The group delay function behaves differently from that of the magnitude spectrum in that it has additive and therefore a higher resolving capability, in comparison to that of the magnitude spectrum.

Traditionally, the phase spectrum of the signal has been ignored, primarily because only the principal values of the phase can be estimated from the Fourier transform. For the phase to be used, the phase function will have to be unwrapped to produce a continuous estimate [3]. On the other hand, the group delay function [4] (defined as the negative derivative of the phase function), which has properties similar to the phase, can be computed directly from the signal. In Section 2 we introduce group delay functions and list some of their properties. The group delay function was successfully used earlier for formant extraction from speech [5], signal reconstruction [6], and spectrum estimation [1]. Another effort on recognition [7] uses the minimum phase group delay function derived from the magnitude spectrum for recognition. In this paper, an attempt is made to extract features directly from the phase function for speech recognition in noisy environments.

The group delay function is undefined if the roots of the transfer function (poles or zeroes) are on the unit circle. Since speech is the output of a stable system, roots that are close to the unit circle are only zeroes. These zeroes are due to either the use of an analysis window to truncate the speech signal or from noise. They are not of much interest in speech analysis, since in speech, zeroes occur only in the production of nasals.

We have defined a new function called the *modified group delay function* which suppresses the zeroes that are close to the unit circle. This is described in Section 3. The modified group delay function is then converted to cepstral features so that it can be used in recognition. In Section 4 we apply the modified group delay cepstrum for recognition of phonemes in a noisy environment [8]. The results are compared to the mel frequency cepstral coefficients (MFCC) and linear frequency cepstral coefficients (LFC) [2].

2. PROPERTIES OF THE GROUP DELAY FUNCTION

Given a discrete-time real signal x(n), the z transform is given by

$$X(z) = \sum_n x(n) z^{-n} \tag{1}$$

We can write X(z) as

$$X(z) = \Pi_i X_i(z) \tag{2}$$

where $X_i(z)$ is either a first-order or a second-order polynomial with real coefficients. The roots of $X_i(z)$ are either real or a complex conjugate pair. The Fourier transform magnitude is a product of the magnitudes of the individual components, and the FT phase is a sum of the phases of the individual components. The group delay function also has properties similar to those of the phase spectrum.

Let

$$X(\omega) = |X(\omega)| e^{j\theta(\omega)} \tag{3}$$

$$log X(\omega) = log(|X(\omega)|) + j\theta(\omega) \tag{4}$$

$$\tau_p(\omega) = -\frac{d\theta(\omega)}{d\omega} \tag{5}$$

where $\tau_p(\omega)$ is the group delay function and can be computed directly from the signal:

$$\tau_p(\omega) = -(\frac{d(log(X(\omega)))}{d\omega})_I \quad (6)$$

$$= \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \quad (7)$$

where the subscripts R and I denote the real and imaginary parts. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x(n)$ and $nx(n)$, respectively. Fig. 1 shows a segment of speech taken from the phoneme *aa* of the Switchboard corpus [9] and its corresponding magnitude and group delay spectra. The first two formants are clearly visible in the magnitude spectrum while the group delay function does not show any structure. This occurs primarily because the segment of speech is a nonminimum phase signal primarily due to the zeroes from windowing and noise. The spikes generated by noise and window effects completely mask the formants corresponding to those of the speech segment. Clearly, what is required is a modification to the group delay function that will yield spectra similar to those of the magnitude spectrum. This is achieved in the modified group delay function, which is described below.

3. THE MODIFIED GROUP DELAY FUNCTION AND FEATURE EXTRACTION

As mentioned earlier, although the group delay function has additive and high resolution properties, for nonminimum phase signals it is ill behaved, especially when the roots are close to the unit circle in the z-domain. In general, for the group delay function to be a meaningful representation, it is only necessary that the roots of the transfer function are not close to the unit circle in the *z domain*. Normally, in the context of speech, the poles of the transfer function are well within the unit circle. The zeroes in speech that correspond to those of nasals also are either within or outside the unit circle since the zeroes also have a finite bandwidth. The spiky nature of the group delay function is primarily caused by pitch peaks, noise, and window effects. In this Section, we modify the computation of the group delay function to suppress these effects. A similar approach was taken in an earlier paper by one of the authors [1] for spectrum estimation.

3.1. Computation of the Modified Group Delay Function

Let $s(n)$ be the clean speech and $x(n)$ the channel and noise corrupted speech. Let $h(n)$ be the time invariant channel response and $w(n)$ the additive noise.

Then we have

$$x(n) = s(n) * h(n) + w(n) \quad (8)$$

$$X(\omega) = S(\omega)H(\omega) + W(\omega) \quad (9)$$

Assuming a source system model for speech production

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Ge(n) \quad (10)$$

$$X(\omega) = \frac{GE(\omega)H(\omega) + A(\omega)W(\omega)}{A(\omega)} \quad (11)$$

$$\tau_x(\omega) = \tau_{numerator}(\omega) - \tau_a(\omega) \quad (12)$$

where $\tau_{numerator}(\omega)$ is the group delay function corresponding to that of $GE(\omega)H(\omega) + A(\omega)W(\omega)$. In regions with high signal to noise ratio, the first term in equation 11 dominates the group delay spectrum in equation 12, while in the regions with low signal to noise ratio, the second term dominates. In either case, the group delay spectrum is dominated by spectral zeroes, in the former due to $E(\omega)$ and in the latter due to $W(\omega)$. The first task is therefore to suppress the zeroes. Once the zeroes are suppressed, we get the following approximation:

$$\tau_x(\omega) \approx \tau_h(\omega) - \tau_a(\omega) \quad (13)$$

where $\tau_h(\omega)$ and $\tau_a(\omega)$ correspond to the group delay functions of the channel and vocal tract system, respectively. Given that the task is speech recognition, the picket fence harmonics are not of much interest. Therefore, these peaks can be suppressed. This can be done by replacing $|X(\omega)|^2$ with a smoothed version of the same $((S(\omega))^2$. This is the modified group delay function. A smoothed version of the spectrum can be obtained by simple cepstral smoothing. The last column in Fig.1 shows the modified group delay spectrum for the same segment of speech. The formants are now clearly visible. But, once again, we have the problem that the peaks at the formant locations are very spiky compared to those of the magnitude spectrum. This could hurt performance. To reduce the spiky nature of the spectrum we have introduced two new parameters into the computation of the modified group delay function, namely, α and γ. The new modified group delay function is defined as:

$$= sign. \left| \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{(S(\omega))^{2\gamma}} \right|^{\alpha} \quad (14)$$

where $sign$ is the sign of the original modified group delay function and $S(\omega)$ is the smoothed version of $|X(\omega)|$. The following is the algorithm for computing the modified group delay function:

1. Let $x(n)$ be the given sequence.
2. Compute the discrete fourier transform (DFT) of $x(n)$ and of $nx(n)$. Let these be $X(k)$ and $Y(k)$, respectively.
3. Compute the cepstrally smoothed spectra of $|X(k)|$. Let this be $S(k)$ [1].
4. Compute the modified group delay function as

$$\tau_x(k) = sign. \left| \frac{X_R(k)Y_R(k) + Y_I(k)X_I(k)}{S(k)^{2\gamma}} \right|^{\alpha} \quad (15)$$

   where $sign$ is given by the sign of $\frac{X_R(k)Y_R(k)+Y_I(k)X_I(k)}{S(k)^2}$
5. Tune the parameters γ and α appropriately for a given environment.

3.2. Conversion of Modified Group Delay Spectra to Meaningful Parameters

The modified group delay function cannot be used directly to train the speech recognition system, since the length of the vector is as long as that of the length of the DFT Window size. To convert the modified group delay function to some meaningful parameters, the

[1] A low order cepstral window (l_w) that essentially captures the dynamic range of $|X(k)|$ is used.

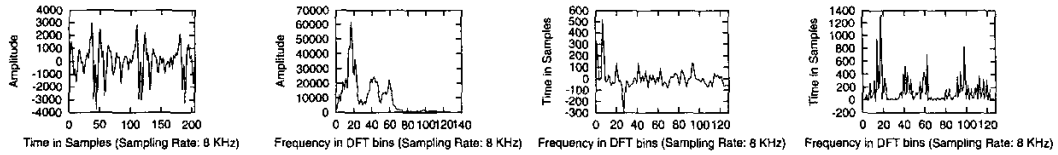

Fig. 1. A Segment of Speech, and Its Corresponding Magnitude, Group Delay, and Modified Group Delay Spectra

group delay function is converted to cepstra using the discrete cosine transform (DCT). Velocity and acceleration parameters for the new group delay cepstrum are defined in the cepstral domain, in a manner similar to that of the velocity and acceleration parameters for MFCC.

3.3. Importance of c_0

In the form of the modified group delay cepstrum defined in Section 3.2, the first coefficient corresponding to nc_n with $n = 0$ is ignored. This value corresponds to the average value in the group delay function. Owing to the effects of (a) linear phase due to the window and (b) the location of pitch peak with respect to the window, it is really not clear how important the first cepstral coefficient is. Further, if we ignore the effects of the window and pitch peak, the group delay must contain additional information, in terms of the delays in sources corresponding to those of the formants. This will result in an average value different from zero in the group delay domain. So, it might not be appropriate to ignore the first coefficient in the inverse DCT. These features are discussed in detail in Section 4.

3.4. Removal of Channel effects in the Group Delay Domain

Owing to the nonlinearities that are introduced by $\alpha \neq 1$ and $\gamma \neq 1$, the removal of the channel effects is an issue. Clearly, from equation 15, if the channel effects are multiplicative, they become additive in the phase and hence the group delay, provided that γ and α each equal 1. Generally, in the case of MFCC, the mean removal is done in the cepstral domain. In the case of the modified group delay function, owing to the artifact introduced by α and γ it is not clear in which domain the mean removal must be performed. We tried two different approaches:

- Ignore the cross terms, and assume that channel effects are additive in the cepstral domain.
- Perform noise removal in the group delay domain with α and γ set to 1.

Although the second approach is theoretically correct, the performance of the system using the first approach seems to be far superior (see Section 4). This could happen because the signal is not only corrupted by multiplicative channel effects but also by additive noise. Using the argument in [1], it is important to suppress the effects of noise in the modified group delay function before it can be further processed.

In the context of the second approach, the mean removal can be performed either on the envelope of the modified group delay function or on the standard modified group delay function. To enable this, a new parameter l_w is introduced. This parameter defines the coarseness of the envelope of the modified group delay function for mean computation.

4. PERFORMANCE EVALUATION

The performance of the system using the new feature is tested on the SPINE database [8]. Phonemes from 15,000 utterances of the SPINE2000 database [8] are used to train the models. Phonemes from 10,000 utterances are used to test the models. Simple isolated Gaussian mixture models (GMMs) with 64 gaussians are trained for each of the 48 phones (including pause) that is present in the database. Two different feature vectors, modgroupdelay_cepstrum_1 (MODGD_1) and modgroupdelay_cepstrum_2 (MODGD_2), are used to refer to the features corresponding to those of the two mean removal mechanisms suggested in Section 3.4. Two more features modgroupdelay_cepstrum1NcN (MODGD_1NcN) modgroupdelay_cepstrum2NcN (MODGD_2NcN) are also defined. They correspond to MODGD_1 and MODGD_2 respectively, and include the average value of the corresponding groupdelay function (see Section 3.3).

Owing to a lack of theoretical insight into the feature, a line search is performed to determine the best front-end for the SPINE database. The experiments performed are listed in Table.1. In the

Experiments
$N_c = 10, 12, 14, 16, 18, 20$
$\gamma = \{0.5 - 1.0\}$ in steps of 0.1
$\alpha = \{0.5 - 1.0\}$ in steps of 0.1
$s_w = 4, 6, 8, 10, 12$
$l_w = 12, 20, 128.$

Table 1. Experiments on front-end parameters for *modified_groupdelay_cepstrum*

experiments it was observed that MODGD_1NcN gives the best results. These results are compared with MFCC and LFC in Table. 2. A feature called GDC, the cepstra derived from the standard group delay function is also included in the Table. It is clear that the performance of the cepstra derived from standard group delay compares very poorly with that of the modified group delay cepstra. Velocity and acceleration are required for speech recognition. Similar to the feature derived from MFCC, velocity and acceleration parameters are defined for the features derived from modified group delay. A composite feature vector consisting of the modified group delay cepstrum, its velocity, acceleration parameters, energy, delta energy, and, delta delta energy is used. The performance of the system for different forms of the modified group delay feature is shown in Table. 3. It is interesting that although the performance of the modified group delay cepstrum alone is far superior to that of MFCC and LFC, the performance of the composite feature is worse. Clearly, the derivative and acceleration parameters are hurting performance, because the nonlinearities in-

featurename	γ	α	s_w	l_w	N_c	% correct
MODGD_1	0.9	0.4	6	128	12	32.85
MODGD_2	0.9	0.3	6	128	12	31.01
MODGD_1NcN	0.9	0.3	6	128	12	39.22
MODGD_2NcN	0.9	0.3	6	128	12	35.61
GDC	1.0	1.0	6	128	12	3.07
MFCC	–	–	–	–	12	34.7
LFC	–	–	–	–	12	35.0

Table 2. Performance of different front-ends on the SPINE database

CompositeFeature	% correct
MODGD_1	**56.24**
MODGD_2	**54.24**
MODGD_1NcN	**57.3**
MODGD_2NcN	**54.88**
MFCC	**60.7**
LFC	**58.24**

Table 3. Comparison of different front-end features

troduced in the modified group delay computation significantly affect the computation of the derivatives. We are currently exploring other approaches to computing these features. Given that the performance of the composite phase feature is worse than the composite MFCC feature, a question now arises: Is there any new information in the phase that is not there in either the composite MFCC or the composite LFC? To answer this, we combine the average of the scores obtained for each of the features individually and recompute the performance. The different combinations and the corresponding performance are given in Table.4. In each of the different combinations, each feature is given the same weight. Performance shows an $\approx$ 2% improvement after the new feature is included. Further, when all three features are combined there is an $\approx$ 2.8% improvement over the best baseline that we have. We have observed that the modified group delay feature performs very well on vowels and pause, while its performance on sounds like *d, l, m, th* and *v* is consistently poor. It is possible that as these sounds have a significant noise component (stops, trill, noise or nasal), zero suppression has resulted in a loss of information. As

Composite Feature	% correct	
MFCC+LFC	**62.54**	
LFC+MODGD_1NcN	**61.67**	
MFCC+MODGD_1NcN	**62.85**	
LFC+MFCC+MODGD_1NcN	**63.62**	Best

Table 4. Combining different front-end features

mentioned earlier, the performance of the modified group delay feature is poor on unvoiced phonemes. This suggests that perhaps the front-end should be different based on the phoneme. It is not yet clear how this can be done.

5. CONCLUSIONS

We have explored the use of features extracted from phase (albeit group delay function) as an alternative or supplement to the features extracted from the magnitude spectrum, namely, MFCC. In the process, we have defined a new feature called the *modified group delay cepstrum*, in which the emphasis is not only on formant locations but also on interformant information. Our experimental results show that the modified group delay function can be used in conjunction with the standard MFCC-based feature in recognition. On a phoneme recognition experiment, the recognition accuracy improved by about 2% absolute over the best baseline MFCC-based system when the modified group delay feature is used in conjunction with MFCC.

Acknowledgments

This research was in part funded by the DARPA EARS program under contract MDA972-02-C-0038. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agency. One of the authors, Hema A. Murthy thanks the Speech Technology and Research Laboratory at SRI International in Menlo Park, California, for supporting her in pursuing this research at SRI International from June through August, 2002.

6. REFERENCES

[1] B. Yegnanarayan and Hema A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Processing*, vol. 40, pp. 2281–2289, September 1992.

[2] V. R. R. Gadde, A. Stolcke, D. Vergyri, J. Zheng, K. Sonmez, and A. Venkataraman, *The SRI SPINE 2001 Evaluation System*. http://elazar.itd.nrl.navy.mil/spine/sri2/presentation/sri2001.html, 2001.

[3] J. Tribolet, "A new phase unwrapping algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, no. 10, pp. 170–177, 1977.

[4] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. NJ: Prentice hall signal processing series, 1998.

[5] B. Yegnanarayana, "Formant extraction from linear prediction phase spectrum," *J. Acoust. Soc. Amer.*, pp. 1638–1640, 1978.

[6] B. Yegnanarayana, D. K. Saikia, and T. R. Krishnan, "Significance of group delay functions in signal reconstruction from spectral magnitude or phase," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 3, pp. 610–622, 1984.

[7] A. Bayya and B. Yegnanarayana, "Noise-invariant representation for speech signals," *Proceedings EUROSPEECH*, vol. 90, pp. 1841–1856, 1999.

[8] Navy Research Laboratory, *Speech in Noisy Environments*. http://elazar.itd.nrl.navy.mil/spine/, 2001.

[9] Linguistic Data Consortium, "Switchboard corpus," 1995.