A NEW GROUP DELAY-BASED FEATURE FOR ROBUST SPEECH RECOGNITION

Erfan Loweimi, *Member*, *IEEE*, and Seyed Mohammad Ahadi, *Senior Member*, *IEEE*

Speech Processing Research Laboratory
Electrical Engineering Department, Amirkabir University of Technology, Hafez Ave., Tehran 15914, Iran
{eloveimi, sma}@aut.ac.ir

ABSTRACT

In this paper we present a novel feature extraction algorithm based on group delay function for robust speech recognition. The modified group delay function (MODGDF) is the main feature extraction method based on group delay function, generally used for robust speech recognition. The recognition tests indicate this feature does not provide notably better results in the presence of additive noise in comparison with MFCC. In the presence of convolutional noise, the performance of MODGDF is considerably lower than MFCC. The method proposed in this paper is simple and makes more efficient utilization of the high resolution property of GDF. It is formed from three main parts which are signal modeling, GDF computation based on extracted model, and compression. The recognition results obtained over AURORA 2.0 task indicate its superior performance in comparison with MODGDF and MFCC.

Index Terms— Robust speech recognition, group delay function, signal modeling, compression

1. INTRODUCTION

It is generally accepted that phase spectrum does not play a significant role in speech processing. The majority of algorithms in this field are only focused on the magnitude spectrum and relatively little attention has been paid to the phase spectrum. In the field of speech enhancement, it is only the magnitude spectrum that is modified. The phase spectrum of the noisy signal is directly transferred to the output with no change, i.e. in the end of the process, enhanced magnitude spectrum is combined with the phase spectrum of the noisy signal and the enhanced signal is synthesized. This is also true for speech recognition, where most of the feature extraction algorithms only utilize the magnitude spectrum and discard the phase spectrum.

The aversion of using phase spectrum in speech processing originated from two main reasons. First, because of phase wrapping, processing and interpreting the phase spectrum become very complicated. It appears that phase-based signal processing is not as mature as magnitude-based signal processing, owing to this phenomenon. The second reason for avoiding phase spectrum is the existence of some well-known perceptual experiments that demonstrate the phase spectrum does not carry a noteworthy deal of intelligibility information [1]-[4]. It is shown that the speech phase spectrum has significant amount of intelligibility information only in such frames as long as one second [3], [4]. However, due to the non-stationarity of speech signal, long frames are not applicable. Hence, there is no attraction for researchers to directly work on the phase spectrum of large frames.

Liu, He, and Palm [5] have conducted a significant research, studying the importance of phase spectrum in speech recognition. Their human-based speech recognition experiments showed that the intelligibility of the phase-only reconstructed speech in frames longer than 128 ms becomes more than that of the magnitude-only reconstructed speech. Alsteris and Paliwal [6], [7] in a similar framework showed that in case of applying a suitable window (rectangular window), even in short frame lengths such as 32 ms, the intelligibility of phase-only reconstructed speech improves and becomes comparable with that of the magnitude-only reconstructed speech.

Shi, Modirshanechi, and Aarabi [8] investigated the importance of the phase spectrum in human speech recognition. They showed that although the role of the phase spectrum in high SNRs is not significant, in low SNRs, it has a remarkable influence on the recognition rate. They combined the magnitude spectrum of noisy signals with the phase spectrum of clean signals and observed that in lower SNRs clean phase spectrum further improves the intelligibility of noisy signal. They claimed that this observation proves the importance of phase spectrum in robust speech recognition. However, it is obvious that the influence of clean signal information inserted in noisy signal, whether of magnitude or phase spectra, increases in lower SNRs and will cause further improvement on the quality and intelligibility of noisy signal. This arises some doubts about their justification in proving the importance of the phase spectrum in robust speech recognition.

As said before, due to phase wrapping, phase spectrum is not directly applicable yet. However, the researchers have

978-1-61284-350-6/11/$26.00 ©2011 IEEE

been trying to use its other representations. Group delay and instantaneous frequency are two major representations of the short-time phase spectrum. Group delay function has found applications in signal reconstruction [9], power spectrum estimation [10] and feature extraction for Automatic Speech Recognition [11], [12]. The modified group delay function (MODGDF) is a group delay-based feature proposed by Murthy and Gadde [11]. However, it does not seem to noticeably outperform MFCC. Particularly in the presence of convolutional noise, it results in notably lower recognition rates in comparison with MFCC. Incidentally, this feature has three parameters (will be discussed in the next section) which should be optimized for maximum recognition rate. It should be noted that line searching over a large data base at four-dimensional space (three parameters as well as recognition rate) for a point which maximizes the recognition rate is a difficult task.

In this paper, we will present a new feature extraction algorithm based on group delay function which results in higher recognition rates in comparison with traditional modified group delay-based feature i.e. MODGDF. In addition, this algorithm only has two parameters that can be simply optimized. The proposed method makes more efficient use of the high resolution property of GDF as well.

This paper is organized as follows. In Section 2 we will briefly review the properties of group delay function and its shortcomings. In Section 3 the proposed algorithm will be introduced. Section 4 deals with the recognition tests results over AURORA 2.0 [13] task as well as their analysis and Section 5 concludes the paper.

2. PROPERTIES OF GROUP DELAY FUNCTION

Group delay function is defined as the negative derivative of continuous phase spectrum. It has two important properties i.e. additive and high resolution [10]. The additive property indicates if two functions convolve with each other in time domain (e.g. a channel impulse response and a signal), they will be added in group delay domain. High resolution property points to sharp peaks of the group delay function, which under specific conditions can be a high-resolution estimation of power spectrum [10].

Although GDF has sharp peaks and consequently high resolution, it is not useful in case of many practical signals in which either zeros or poles get close to the unit circle. In case of speech signals, since zeros are close to the unit circle, the group delay function cannot successfully estimate the power spectrum. In fact, these zeros result in spurious peaks which mask the formants leading to a very poor estimation of power spectrum. To overcome this problem two methods have been proposed. Yegnanarayana and Murthy [10] proposed cepstral smoothing to eliminate the effects of zeros introduced by excitation component of the speech signal. The modified group delay function (MGDF) is a GDF computed through cesptral smoothing [10]. Bozkurt and Couvreur [14] proposed chirp group delay function (CGDF) dealing with the aforementioned problem. This method involves two stages. First, zeros located outside the unit circle must be eliminated. This will reduce the speech signal to its minimum phase form. Then, the *z*-transform is evaluated on a circle whose radius is greater than unity. The proposed radius in [14] is 1.12. This method has high computational overhead required to extract the zeros of the signal.

Murthy and Gadde [11] proposed a feature based on modified group delay function. In this method, they further modified the modified group delay function in following form

$$\tau_x(k) = sign.\left|\frac{X_R(k)Y_R(k)+X_I(k)Y_I(k)}{S(k)^{2\gamma}}\right|^{\alpha} \quad (1)$$

where $\tau_x(k)$ is a new modified group delay function of *x(n)*, the subscripts *R* and *I* denote real and imaginary parts, *X(k)*, *Y(k)* and *S(k)* indicate the Fourier transform of *x(n)*, *nx(n)* and cepstrally smoothed spectrum of $|X(k)|$, respectively, and *sign* is the sign of $\frac{X_R(k)Y_R(k)+X_I(k)Y_I(k)}{S(k)^2}$. In the next stage, by taking Discrete Cosine Transform (DCT) the features (MODGDF) will be extracted from $\tau_x(k)$.

As seen, there are 3 parameters, α, γ, and s_w that should be optimized for maximizing the recognition rate. s_w is the window length in cepstral domain, used to remove the effect of the excitation component of speech signal which is the main reason of introducing zeros close to the unit circle. These zeros make the group delayed-based estimated power spectrum spiky and consequently worthless. α and γ are used to adjust the bandwidth and sharpness of the peaks (formants) as well as compression. In [12] the proposed value for α, γ, and s_w are 0.3~0.4, 0.9, and 4~9, respectively. However, the question which arises here is that whether these values for aforementioned parameters lead to global maximum of the recognition rate in any database. In addition, do these values still remain optimum choices in the presence of additive and convolutional noises? Actually, line searching in 4-dimentional space (α, γ, s_w, and recognition rate) is not an easy task. Furthermore, if an optimum point was found, there is no guarantee that this optimal choice, over different databases or in the presence of different noise types or SNRs, remains the same. Due to lacking theoretical insight into this feature, the aforementioned problem becomes more complicated.

The recognition tests showed that the MODGDF features do not provide notably better results in comparison with MFCC [11], [12]. Moreover, as will be shown in Table 1, in the presence of convolutional distortion, the performance of MODGDF is considerably lower than MFCC.

3. THE PROPOSED METHOD

In the previous section we discussed some of the shortcomings of MODGDF feature. In this section, we

present our method, which does not suffer from the problems of MODGDF and leads to noticeably better recognition rates in comparison with MFCC in the presence of both additive and convolutional distortions.

Figure 1 shows the block diagram of the proposed method. This algorithm consists of three main parts, i.e. signal modeling, GDF computation and compression. The speech signal does not pass the pre-emphasis block. After frame blocking and windowing, an Autoregressive (AR) model is extracted for each frame. In this stage, we have used LPC and Burg [16] methods. Then, based on the extracted model, the GDF is computed and subsequently, compressed into k_2 elements through two-stage DCT.

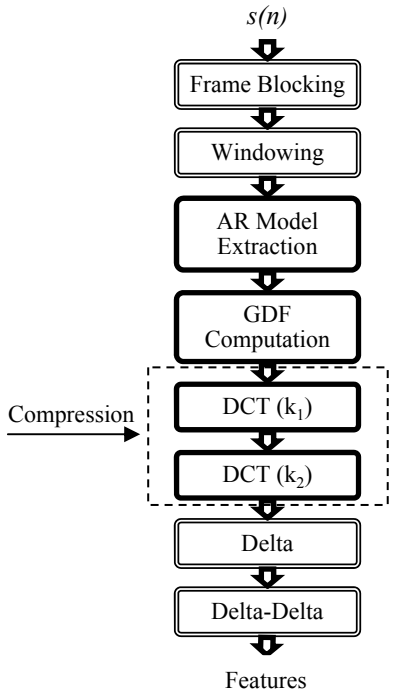

Fig. 1. Block diagram of the proposed method.

Coupling the GDF with AR model provides the following advantages:

1) The model will be AR; therefore, the effects of zeros which cause a noise-like GDF and mask the formants are highly alleviated.
2) The bandwidth and sharpness of the formants will be adjusted by the use of an appropriate order for the AR model.

As a result, this coupling provides more efficient utilization of the high resolution property of GDF in the power spectrum estimation. Figure 2 shows the GDF, modified GDF (MGDF), LPC and Burg power spectra, and GDFs calculated based on AR extracted models for a typical speech signal.

Next step is compression of the GDF into a vector with 12 (11~13) elements. Compressing the samples of a frame which typically has 256 or 512 elements into 12 samples by taking one-stage DCT drastically increases the compression loss. It is another problem of MODGDF feature that should be taken into consideration. Dealing with this issue, we do the compression in two stages. At first we compress the samples into k_1 elements. Then, we compress these numbers of samples into k_2 elements. It is clear that the value of k_2 is around 12. For this, we must find the proper value for k_1. To do this, we follow some guidelines from the MFCC feature extraction procedure. In MFCC, the samples of the power spectrum of each frame are first compressed as the energies of outputs of a number of filters (say 23). Accordingly, one may suppose the suitable value for k_1 must be within this range. In addition, we expect the appropriate value for k_1 to be larger than the number of filters in MFCC. The reason is that, in this case, there is no emphasis on a specific range of frequencies like that of mel-filter bank. Our simulations show that 30 is a suitable choice for k_1. However, it is not a critical decision and other values in this range can be used.

It should be mentioned that this type of compression does not work for compressing the MODGDF and magnitude spectrum. It seems that the smooth structure of the GDF computed based on AR model makes this method work. Besides, we observed that compressing the power spectrum of AR model through taking two-stage DCT results in lower recognition rate in comparison with compressing the GDF. It is due to high resolution property of GDF that is employed more efficiently in this algorithm.

There are other points that should be investigated. In [15], we have showed that Chebyshev window with dynamic range of 30 dB is almost the best choice for working with the phase spectrum. Phase-only reconstructed speech along with this window has the highest quality. Contrary to MODGDF, we observed that for the proposed method applying this window results in higher recognition rates, in comparison with the Hamming window.

The last points that should be discussed here are the AR modeling method and its order. The proper order for an 8 kHz-sampled speech is in the range of 8 to 12. We found 12 a better choice, although like k_1 it is not a critical choice. Here, we have used LPC and Burg methods [16] for AR modeling. LPC method determines the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense. It uses the autocorrelation method of autoregressive modeling to find the parameters. The Burg method estimates the reflection coefficients and uses them to estimate the AR coefficients recursively. The results show that Burg method leads to better recognition results, in comparison with LPC method. It should be noted that the computational load of Burg method is higher than LPC. So, the LPC could be considered as a more economic choice.

We name the proposed method ARGDD because of AR model extraction, GD computation and Double DCT operation for compression. ARGDD1 and ARGDD2 in Table 1 and Figure 3 refer to employment of LPC and Burg methods, respectively.

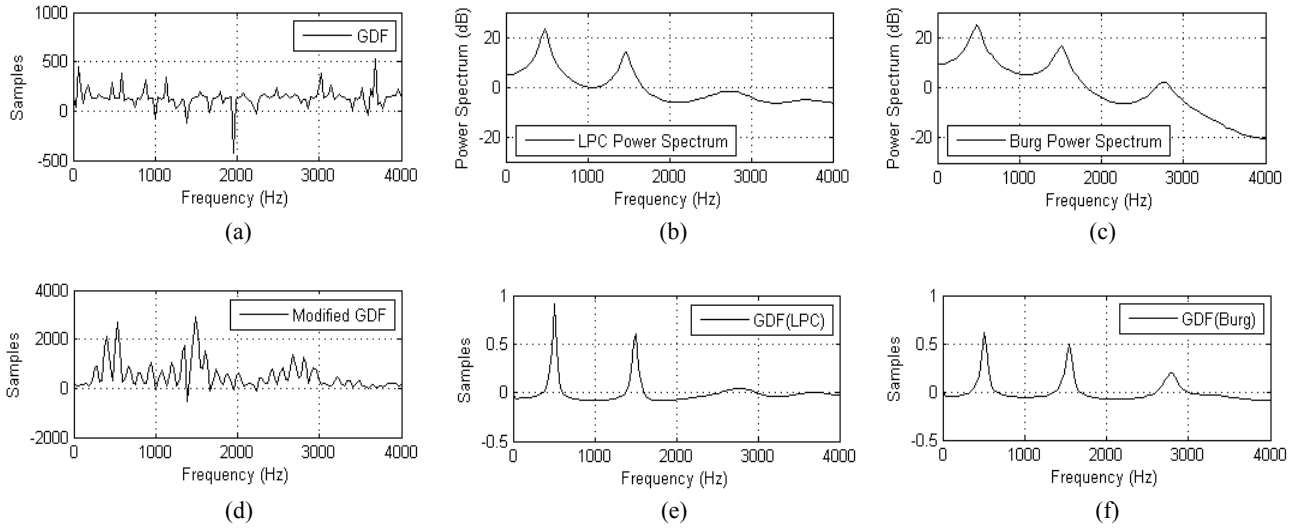

Fig. 2. (a) Group delay function, (b) LPC power spectrum, (c) Burg power spectrum, (d) modified group delay function, (e) GDF (LPC), (f) GDF (Burg).

4. RECOGNITION TESTS RESULTS AND THEIR ANALYSIS

In all cases, the feature vector consists of 39 elements containing static coefficients which include C_0 or log-energy as well as dynamic and acceleration coefficients. In MFCC computation, the pre-emphasis coefficient is 0.97 and the number of filters of mel-filter bank is 23, in the range of 64 to 4000 Hz. For all the presented features, frame length and frame shift are set to 32 and 12 ms, respectively. Hamming window has been used in case of MFCC and MODGDF. For the proposed method, Chebyshev window with dynamic range of 30 dB have been used. We used HTK [17] to train and test the HMMs. In all of the features tested here, cepstral mean normalization (CMN) has been performed.

As seen in Table 1 and Figure 3, the ARGDD results in interesting recognition rates over clean-trained AURORA 2.0 task [12] in the presence of both additive (A and B test sets) and convolutional (C test set) distortions. The performance of our method is about 10% (absolute) or more above MFCC performance in SNRs of 5 dB and less on average. It is also has a notably higher performance in case of convolutional noise (C test set) compared with MFCC and especially MODGDF. Due to similarity of ARGDD1 and ARGDD2 recognition results and trends, we just depicted the results of ARGDD2 in Figure 3 to increase the visibility and avoid cluttering.

As it is shown in [11] and [12], for MODGDF, augmenting the feature vector with C_0 will increase the recognition rate. However, in case of MFCC, augmenting the feature vector with logarithm of energy (log-energy) is a better choice. Table 1 shows that using C_0 is a relatively suitable choice for the proposed feature. However, we observed that the centralized log-energy along with ARGDD results in higher recognition rates. As seen in Table 1, using centralized log-energy in case of ARGDD and MODGDF, results in higher recognition rate. It is not yet clear why this happenes. However, taking a look on the contradictory effect of centralizing the C_0 in MFCC and MODGDF shows that the role and influence of appending the feature vector with energy or C_0 is not identical among different front-ends. In case of MODGDF, if the CMN includes C_0, the recognition result will be higher, but for MFCC, excluding C_0 from CMN leads to better recognition rate. In a similar manner, centralizing the log-energy, in case of MFCC feature decreases the recognition rate while in case of MODGDF and ARGDD improves the performance. As well, we observed that excluding C_0 from CMN, similar with MFCC, results in higher recognition rate in case of ARGDD.

Table 1. Average (0-20dB) word accuracy as percentage for the Aurora 2.0 Task

	Test Set A	Test Set B	Test Set C
MFCC-E	69.38	72.88	69.95
MFCC-E*	65.30	70.35	63.55
MFCC-C_0(C_0^+)	63.97	67.87	64.16
MFCC-C_0(C_0^-)	64.36	66.97	70.73
MODGDF-E	65.35	70.17	53.56
MODGDF-E*	68.21	72.93	56.81
MODGDF-C_0(C_0^+)	68.15	72.48	54.90
MODGDF-C_0(C_0^-)	59.02	66.18	48.75
ARGDD1-C_0(C_0^-)	69.68	69.13	70.82
ARGDD2-C_0(C_0^-)	68.46	71.40	**73.86**
ARGDD1-E*	**74.87**	75.84	69.68
ARGDD2-E*	74.58	**77.42**	73.18

E* : Centralized log-energy
C_0^+ : CMN included C_0
C_0^- : C_0 excluded from CMN

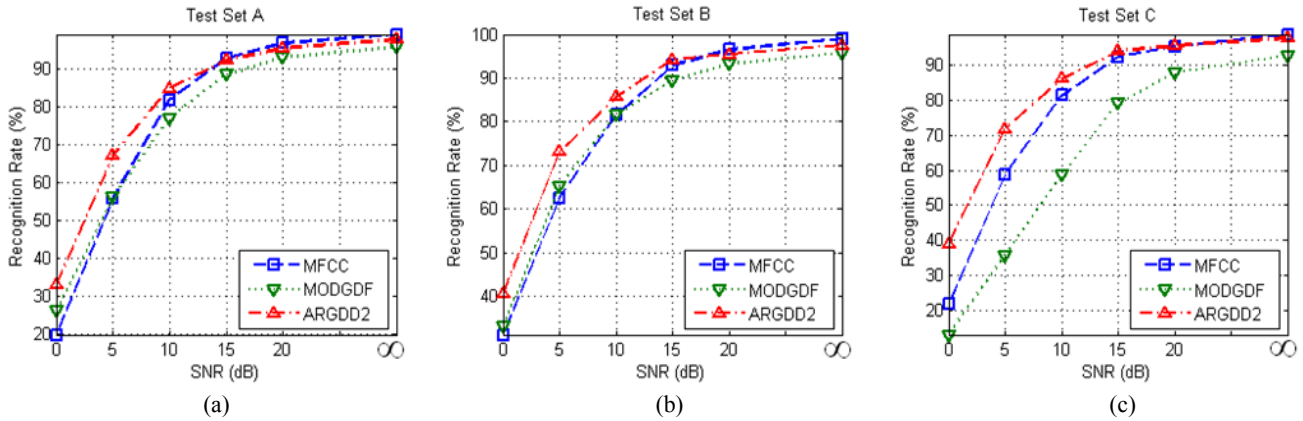

Fig. 3. The recognition rates of the ARGDD2-E*, MODGDF-E*, and MFCC-E versus SNR on the AURORA 2.0 task (the best cases for each feature based on Table 1). (a) Test set A (includes subway, babble, car, and exhibition additive noises), (b) Test set B (includes restaurant, street, airport, and train-station additive noises), (c) Test set C (includes subway and street convolutional noises).

5. CONCLUSION

In this paper we presented a novel feature extraction algorithm based on group delay function for robust speech recognition. The proposed algorithm, ARGDD, consists of three main parts, those are, AR signal modeling, GDF computation based on extracted model, and compression through two-stage DCT. This method, in comparison with the MFCC and MODGDF, has notably higher performance in the presence of both additive and convolutional noises. Its high recognition rates on AURORA 2 task shows the noteworthy potentials of group delay function and phase spectrum to be used in speech recognition. Decreasing the loss of compression through applying more efficient methods and embedding other blocks which could increase the recognition rate are two avenues for further exploration and future works.

6. REFERENCES

[1] G. S. Ohm, "Uber die Definition des Tones, nebst daran geknupfter Theorie der Sirene und ahnlicher tonbildender Vorrichtungen," *Ann. Phys. Chem.*, pp. 513–565, 1843.

[2] H. L. F. von Helmholtz, "On the sensations of tone" (English translation by A.J. Ellis), Longmans, Green and Co., London, 1912 (original work published 1875).

[3] A. V. Oppenheim, J. S. Lim, G. E. Kopec, and S. C. Pohlig, "Phase in speech and pictures," in *Proc. IEEE Int. Conf. Acoust., Speech and SignalProcesring,* pp. 632-637, Apr. 1979.

[4] A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," *Proc. IEEE*, vol. 69, pp. 529-550, May 1981.

[5] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, pp. 403-417, 1997.

[6] K. K. Paliwal and L. D. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. of Eurospeech-2003*, pp. 2117–2120, 2003.

[7] L. D. Alsteris and K. K. Paliwal, "Importance of window shape for phase-only reconstruction of speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Processing*, pp. 573–576, 2004.

[8] G. Shi, M. Modirshanechi, and P.Aarabi, "On the importance of phase in human speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1867-1874, Sep. 2006.

[9] 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 Process.*, vol. ASSP-32, pp. 610–622, Jun. 1984.

[10] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Process.*, vol. 40, pp. 2281–2289, Sep. 1992.

[11] H. A. Murthy and V. R. R. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process.*, Hong Kong, China, vol. I, pp. 68–71, Apr. 2003.

[12] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech and Language Proc.*, vol.15, pp. 190–202, Jan. 2007.

[13] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition Systems under noisy conditions," in *Proc. ASR2000*, Paris, France, Sep. 2000.

[14] B. Bozkurt and L. Couvreur, "On the use of phase information for speech recognition," *European Signal Processing Conf.*, 2005.

[15] E. Loweimi, S. M. Ahadi, and H. Sheikhzadeh, "Phase-only speech reconstruction using short frames," *IEEE Trans. Audio, Speech and Language Processing,* submitted for publication.

[16] S. M. Kay, "Modern spectral estimation: theory and application," Englewood Cliffs, NJ: Prentice Hall, 1988, pp. 228–230.

[17] S. Young, "The HTK book," Cambridge University Engineering Department, Cambridge, UK, 2001.