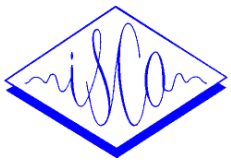

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

FILTERING OF SPECTRAL PARAMETERS FOR SPEECH RECOGNITION

Climent Nadeu and Biing-Hwang Juang*

AT&T Bell Laboratories, Murray Hill, NJ 07974, USA

E-mail: nadeu@tsc.upc.es

ABSTRACT

The time sequences of speech parameters resulting from current short-time spectral estimators show a tradeoff between estimation error variance and time and frequency resolution. In this paper, we apply frequency analysis and linear filtering to these sequences to gain insights into their limitations and to provide an interpretation framework for several parameter processing techniques proposed in the past. Particularly, the observation of their long-term spectrum reveals the importance of band equalization for improving discrimination in speech recognition. Based on that, we propose a method of filtering the sequences that includes an explicit equalization and incorporates a bandwidth parameter. By using Slepian sequences in the design of the filters, good results were obtained in our preliminary word recognition tests.

1. INTRODUCTION

The first step in the pattern matching approach to the problem of speech recognition is to convert a speech waveform into a sequence of features, usually in the form of spectral parameters [1]. Speech signals are usually modeled as the output of a time-varying filter driven by a signal whose spectrum is essentially either flat or a train of spectral lines of equal power. Consequently, on a short-time basis, the envelope of the speech spectrum represents the instantaneous spectral response of the filter whose characteristics are the determining factor of the identity of a speech sound or a speech utterance. Conventionally, speech spectral envelopes are represented by means of all-pole models or various forms of periodogram averaging, and often are expressed in terms of the corresponding cepstral coefficients [1].

These representations are calculated via short-time spectral analysis. Let the sampled speech signal be $s(l)$. A window function $w(l)$ is applied to it at regular intervals nN_0, n=...,-1,0,1,2,... to form frames of windowed signal $s(l)w(nN_0-l)$. The window function is usually of finite duration L_0. Spectral analysis techniques are then used to obtain a short-time spectral estimate for each signal frame which is represented with Q parameters (Q is the order of the all-pole model, or the number of frequency bands of the periodogram-based estimators). Thus the process results in a set of time sequences of spectral parameters that represent the temporal evolution of the spectral response of the time-varying filter. We shall refer to each time sequence of spectral parameters as TSSP.

There are certain inherent limitations in this type of speech signal representation. First, spectral estimation based on finite data involves a certain random estimation error. Moreover, in speech spectral estimation, the relative positioning of each frame with respect to pitch periods introduces an additional estimation error. Given a spectral estimator and a number of parameters Q, the estimation error variance strongly depends on the window length. Asymptotically, the relative variance is inversely proportional to the window length in the above mentioned conventional representations. However, increasing the window length, while may lead to a reduced estimation variance for steady state sounds, will cause loss of temporal resolution. This is due to the fact that the change of speech characteristics does not synchronize with the fixed sampling interval, N_0, of data frames and that a longer data window would be more likely to encompass transients as well as different sounds in one frame.

It is not advisable to unlimitedly reduce the frame sampling interval N_0 either to gain time resolution. Note that human's articulatory apparatus does not move arbitrarily fast. If we view the set of TSSPs as a process that characterizes the articulatory movements, the process will not have an infinite frequency range. For a given window length L_0, increasing the frame rate (i.e. reducing the framing interval N_0) would not necessarily result in more time resolution, due to the higher correlation that would be obtained between consecutive TSSP samples. If the window length L_0 were reduced as well, more error would be included in the estimate.

Thus, there is a tradeoff between error variance and time resolution of the spectral estimator. Analogously, there is a tradeoff between error variance and frequency resolution. A higher frequency resolution may allow a better representation of the fine structure of the spectral envelope. For a given window length L_0, the number of spectral parameters Q determines that tradeoff for each estimator.

The purpose of the paper is to apply frequency analysis and linear filtering theory to the TSSP to gain insights into the speech feature extraction or estimation process, and to propose a filtering method for changing tradeoffs between the above mentioned estimation performance factors, particularly in speech recognition applications.

2. SPECTRUM OF TSSP

Let $logS(\omega,n)$ be the short-time log spectral estimate of the speech signal with n denoting the frame index and ω the frequency. We shall use cepstrum $c(m,n)$ as the representation of $logS(\omega,n)$, i.e.

$$c(m,n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log S(\omega,n) e^{j\omega m} d\omega \qquad (1)$$

due to its widespread use in speech recognition applications. Note that the Fourier transform of the time sequence of the m th cepstral coefficient $c(m,n)$ is

$$C(m,\theta) = \sum_n c(m,n) e^{-jn\theta} \qquad (2)$$

where the frequency variable θ is the counterpart of the frame index n.

The long-term power spectrum of the TSSP will be denoted by $T(m,\theta)$. In the current case, it can be empirically estimated by computing and averaging

$$|C(m,\theta)|^2$$

* C. Nadeu is with Universitat Politècnica de Catalunya, Barcelona, Spain. This work was carried out during his sabbatical leave, while he was a visiting researcher at Bell Laboratories.

over a data base of speech signals. Fig.1 shows a typical long-term power spectrum $T(m,\theta)$ as a function of θ (in Hz) obtained by averaging over a large set of isolated digit utterances and speakers, and over the first 12 LPC-cepstral coeficients m=1,2,...,12 (calculated like in section 5).

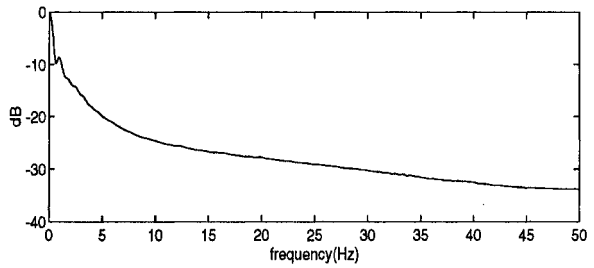

Fig.1 Typical TSSP long-term power spectrum

While the long-term power spectrum of each m-th TSSP obviously can be a function of the utterances, the type of spectral parameters, the speakers, etc. used in averaging, it generally decreases along the θ axis and can be reasonably well approximated by the spectral response of the first-order filter

$$F(z)=\frac{1}{1-\rho z^{-1}} \tag{3}$$

where ρ is close to one.

These long-term power spectra, as argued previously, contain noisy components which are due to the estimation errors. One way to measure the estimation error is to perform an experiment as follows. A sequence of unit variance white noise was used as the input to a time-invariant filter $H(z)$, the frequency response of which was chosen from a typical speech spectrum. The output of the filter then was LPC analyzed with a window length L_0=240 and a frame shift N_0=80 (these conditions were identical to those used in creating Fig.1). The estimated spectral sequence was further transformed to the cepstral sequence, of which we only examined the first 12 coefficients, resulting in 12 TSSP $c_S(m,n)$, m=1,...12. For an unbiased estimator, the long-term power spectrum of a TSSP of a stationary signal for $\theta\neq0$ is a measure of the estimation error variance in the frequency domain. Ideally, in absence of estimation error, the cepstral sequence would be constant, i.e.

$$c_s(m,n)=\bar{c}_s(m)\xleftrightarrow{FT}\log|H(\omega)|^2 \tag{4}$$

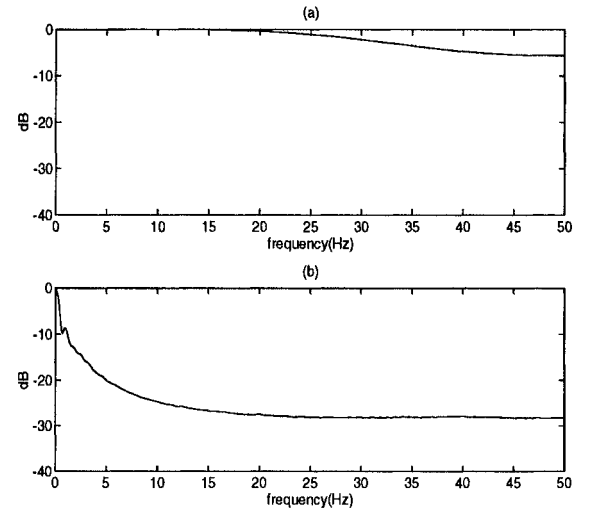

Fig.2 (a) Estimated spectrum of the cepstral error sequence. (b) Ratio between the spectrum of Fig.1 and the cepstral error spectrum.

Fig.2(a) is a plot of the estimated power spectrum of the cepstral error sequence resulting from subtracting the average value to $c_S(m,n)$ (8KHz sampling rate is assumed). Note that since the analysis window is larger than the frame shift interval, i.e. $L_0>N_0$, the overlapping of adjacent data frames produces a certain degree of correlation, resulting in a non-flat spectrum as can be seen in the figure.

The error spectrum can be directly compared to the spectrum of TSSP to gain insight into the reliable bandwidth of the TSSP estimates. Fig.2(b) shows the ratio between the TSSP spectrum of Fig.1 and the estimation error spectrum. Notice from the plot that the ratio (which resembles a signal-to-noise ratio) is essentially flat beyond a frequency θ_S which is around 15 Hz in this case. It is thus reasonable to assume that beyond θ_S, the frequency content of the TSSP of a speech signal contains a significant amount of estimation error and thus bears unreliable information.

3. FILTERING OF TSSP

Dynamic features of speech in the form of differential parameters are extensively employed in speech and speaker recognition systems. The differential parameters are usually analyzed in the time domain, as successive derivatives that capture the change of the TSSP [2] (Taylor's expansion). However, they can also be envisioned as the output of a linear filter driven by the TSSP. In this sense, these parameters can be referred to as *filtered parameters.*

Probably the most common version of filtered parameter that approximates the derivative is the usually called regression coefficient or delta-cepstrum [2]. Its associated impulse response is the first degree discrete Legendre polynomial. Another usual version is the one that computes the difference of two non-consecutive TSSP samples (difference filter) [3]. Higher differential parameters are usually calculated by using higher degree Legendre polynomials [4] or by applying the difference filter to the first differential parameter (as in [5]).

On the other hand, the TSSP has also been filtered to remove or attenuate its zero and very low frequency components when they are contaminated by quasi-invariant linear distortion of the speech signal [6]. Thus we might say that so far the filtered parameters have been intended: 1) as a supplement of the unfiltered parameter for improving the discrimination capability, and 2) as a substitute of it for removing the linear distortion.

It can easily be shown that each of those filters that compute supplementary features has two basic components: 1) a differentiation component that corresponds to a zero at z=1 (θ=0), and 2) a smoothing component that bounds the pass-band. This is also often so for filters that compute substitutive parameters. For instance, the IIR filter used in [6] essentially consists of that zero at z=1 plus a pole whose magnitude is close to one, which controls the pass-band cutoff frequencies.

3.1 Spectral effects of filtering

On the one hand, according to expression (3), the spectrum of the TSSP $T(m,\theta)$ can be approximately equalized by filtering the TSSP with a first-order FIR filter showing a zero at $z=\rho$. On the other hand, we pointed out in the preceeding section that filters applied to the TSSP usually show a zero at z=1. Thus, since ρ is close to one, the differentiation component of those filters is actually equalizing the spectrum of the TSSP, except the very low frequency region, where the spectral components are largely attenuated and the zero frequency component is removed. Viewing $c(m,n)$ as a stationary random process for each m, that equalization implies to uncorrelate the TSSP.

The part of the above mentioned filters that would result from excluding the zero at z=1 selects the frequency band where the power of the filtered TSSP is most concentrated. We have observed in our investigation that, when several filtered parameters are used together in speech or speaker recognition, their spectral bands are separated, distributed in a frequency interval $0\leq\theta\leq\theta_C$, where θ_C may depend on the

recognition system and the recognition task. Actually, both Legendre and cascaded difference filters lead to distributed pass-bands since they include an additional zero at z=1 each time their degree increases by one. According to the observation of section 2 regarding the existence of a frequency θ_S beyond which the TSSP spectrum essentially consists of estimation error, the spectral bands of the filtered parameters should not exceed θ_S, i.e. $\theta_C \leq \theta_S$. In fact, the results of an experiment reported in [7] with a filter that selects a band beyond that effective bandwidth were negative. Fig.3 shows the spectrum of the cepstral sequence corresponding to the DARPA Resource Management (RM) FEB89 test data base along with the output spectra of the two filtered features used in [5], i.e. the first Legendre feature with L=5, and the feature resulting from applying the same filter with L=3 to the first one. Note that the bandwidth θ_C of the unfiltered sequence is around 20-25 Hz, larger than for the digit spectrum in Fig.1.

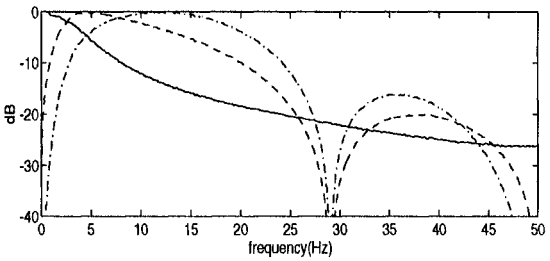

Fig.3. RM spectrum (–), and spectra of the first (--) and second (.-) filtered features [5].

3.2 Tradeoff between resolution and estimation error power

On the one hand, the equalization performed by the filter produces an augmentation of the time resolution (in the sense of a dynamics amplification) of the TSSP. On the other hand, the band-pass characteristics of the filter attenuates the likely unreliable components of the TSSP spectrum located at higher frequencies. Consequently, we can interpret the filtering of TSSP as changing its tradeoff between time resolution and estimation error power. Indeed, each filtered parameter shows its own trade-off.

There exists an interesting analogy between filtering of TSSP and the cepstral liftering commonly used for distance computation in some recognition systems [1]. Actually, cepstral liftering is equivalent to a filtering process on $logS(\omega,n)$ along the variable ω; furthermore, it performs a variance (power) equalization [8] in a certain cepstral band, and a removal of both the (unreliable) high quefrency and zero quefrency components. Consequently, it changes the tradeoff between frequency resolution and estimation error power.

However, unlike what occurs in the quefrency domain, where the power of the estimation error decreases along the quefrency index [9], the power of the estimation error in the time domain n is essentially constant in the region $0 \leq \theta \leq \theta_S$ (assuming it is uncorrelated with the TSSP). Consequently, equalization involved in higher order filtered parameters may unduly boost the estimation error at higher frequencies. This fact might explain why filtered features of order higher than 2 are not employed in continous speech recognition systems, where the pass-bands of the features are usually located at higher frequencies than in isolated word recognition systems (for example, typical lengths of filters may be 5 and 7 for continous speech recognition [5], and 9 and 25 for isolated word recognition [4]). Moreover, it can also help to explain why the so-called delta-delta-cepstrum (a second order filtered feature as the one used in Fig.3) may not show a consistent good performance [5]. Observe its relatively high frequency content in Fig.3.

Thus, interpreting the computation of differential parameters and cepstral liftering as filtering processes helps to understand the tradeoff between resolution (both in frequency and time) and estimation error power.

4. FILTER DESIGN

The distribution of the filtered parameter spectral bands along a frequency interval $0 \leq \theta \leq \theta_C \leq \theta_S$ that obtain the best recognition results may depend on several factors: number of supplementary features, type of recognition task (e.g. IWR or CSR), size of speech units, speaking rate, noise characteristics, etc. Consequently, the structure of the filters that compute the supplementary or substitutive parameters should be flexible enough to allow adaptation to those factors. The structure involved in both Legendre and difference filters consists of a successive differentiation plus low-pass smoothing. Most reported differential parameters are obtained either with it or with slight modifications of it.

In order to increase the structure flexibility, we propose to separate spectral equalization from filtering (band splitting), where the filters should be tailored according to the above mentioned factors. Since the optimal filter specifications for a given set of factors are unknown, we might resort to a set of parameterized filter responses that were sufficiently flexible to be empirically tuned to those factors. For example, simple IIR filters with a complex pole controlling the band center and bandwidth might be used.

4.1 Slepian filters

An even simpler, though less flexible, approach consists of assuming orthogonality among filter responses. For a given length L of the impulse responses, there are L orthogonal filters whose pass-bands are distributed in the whole frequency range $0 \leq \theta \leq \pi$. Actually, only the filters whose pass-bands are included in the interval $0 \leq \theta \leq \theta_C$ are needed. Therefore, discrete prolate spheroidal wave (or Slepian) sequences [10] appear good candidates for the desired impulse responses of the filters since: 1) they are also orthogonal in a band $0 \leq \omega \leq W$ ($W=\theta_C$ in the present application) and 2) they show a maximum concentration of power in that band.

Slepian sequences $v_k(n)$ depend on two parameters, their length L and the bandwidth W. They are defined through the eigensystem of equations

$$\sum_{m=0}^{L-1} v_k(m) \frac{\sin W(n-m)}{\pi(n-m)} = \lambda_k v_k(n) \qquad (5)$$

where $0 \leq n \leq L$-1, $0 \leq k \leq L$-1, and λ_k is the fraction of energy that lies inside the band $0 \leq \omega \leq W$. Since

$$\sum_{k=0}^{L-1} \lambda_k = LW/\pi \qquad (6)$$

given the first K Slepian sequences, i.e. $v_k(n)$, k=0,...,K-1, the time-bandwidth product LW controls their λ_k, and thus the secondary lobes of their frequency responses. The quotient L/W almost determines their main lobes as long as λ_k is close to one.

4.2 Filter design procedure

K denoting the number of Slepian filters, we propose the following design procedure:

1. Equalize the TSSP spectrum with $1-rz^{-1}$, where r depends on the training data base. (Actually, only equalization up to $\theta_C \leq \theta_S$ is needed).
2. Choose W and L so that

$$LW/\pi \geq u(K) \qquad (7)$$

where $u(K)$ is positive and such that λ_K is sufficiently close to one to allow that the secondary lobes of the K-th Slepian filter response are lower than a certain value. We used $u(K)=K+1$ in our experiments. If L and W are such that (7) is verified with equality, W approximately coincides with the highest frequency of the main lobe of that filter response. Thus, if θ_C is given, $W=\theta_C$, and $L=\pi u(K)/\theta_C$. Indeed, if the bandwidth θ_C is not known a priori, (7) only gives a constraint on the value of LW, and the quotient L/W has to be chosen to obtain best recognition results for a given recognition task and a given system.

5. EXPERIMENTAL RESULTS

In order to validate both the meaningfulness of the filtered TSSP spectrum and the usefulness of the Slepian filters, we applied the above design method to a speaker-independent word recognition task. Preliminary tests were conducted using: 1) only a filtered set of parameters (one feature), and 2) the unfiltered set and two supplementary filtered sets (three features).

5.1 Data base and speech recognition system

The data base consisted of single digits embedded in silence and was collected using two microphones and through the telephone network. 2198 utterances from 50 speakers were used to train the digit models, and 919 utterances from 25 speakers were used for testing. Speakers were balanced by gender.

A recognizer based on continous density hidden Markov model was used in the tests. Each of the 11 digit models consisted of 10 states, and the silence model had 5 states. Only one diagonal covariance Gaussian pdf was used per state. The 8 KHz sampled speech was pre-emphasized with the filter $1-0.95z^{-1}$ and autocorrelation coefficients were computed every 10 ms, using a 30 ms Hamming window. After a 10-order LPC analysis, 12 cepstral coefficients plus the energy were computed per frame. The time sequence of each coefficient as well as that of the energy were filtered for every utterance.

5.2 One feature (substitution case)

For equalizing, r=0.97 was chosen since for this value the equalized TSSP spectrum of the training data base is approximately flat. In (7), $u(1)=K+1=2$ was chosen after observing that the secondary lobe height of $v_0(n)$ was -23 dB (λ_0 slightly larger than 0.98) for the resulting minimum value $LW=2\pi$.

L	12	13	14	15	16
W=8 Hz	35	29	30	29	31
W=10 Hz	32	26	26	28	29
W=12 Hz	30	27	31	24	26

Table1 Recognition errors using one substitute Slepian feature.

Table1 shows the number of recognition errors for several values of L and W. The number of errors without filtering the cepstral and energy sequences is 62, much larger than using the filter. Note from the table that similar values of L/W lead to similar scores. The best score is obtained for $L/W=62.5/\pi$. Although the case L=12 and $W=0.16\pi$ (8 Hz) does not satisfy (7), it was included in the table for illustration since, even it has a quotient value close to $62.5/\pi$, it shows a high number of errors.

Using r=1 instead of r=0.97, the number of errors substantially increased. This result suggests that a complete cancellation of the zero frequency component is not desirable. Hence, the strong error reduction produced by the filter seems due not only to the attenuation of the zero frequency but also to the spectral equalization performed in the low frequency region. Empirically optimizing the value of the real pole of a filter as in [6], 25 errors were obtained with a pole at 0.8. The corresponding filtered TSSP spectrum is quite similar in its high power band to the Slepian filter one for L/W around $62.5/\pi$.

Since $W \approx \theta_c$ if (7) is satisfied with equality, the best recognition results are obtained in this experiment with a bandwidth θ_c around 10 Hz.

5.3 Three features (supplementation case)

In this case K=2. In (7), $u(2)=K+1=3$ was chosen. The secondary lobe height of $v_1(n)$ is -18 dB (λ_0 is 0.97) for the corresponding minimum value $LW=3\pi$.

Table 2 shows the number of recognition errors for several values of L and W, using again r=0.97. Notice that the best scores correspond to $L/W=125/\pi$. Moreover, extending W to 14 and 16 Hz but keeping that optimum L/W value,10 errors were also obtained. Conversely, with that optimum value and W=6 Hz, 15 errors resulted; however, in this case, the product LW is much lower than its selected minimum value 3π.

L	15	20	25	30	35
W=8 Hz	15	11	14	16	17
W=10 Hz	15	13	10	14	15
W=12 Hz	15	15	15	11	14

Table 2. Recognition errors using three features and Slepian filters.

Using r=1 instead of r=0.97, the number of errors increased like for the one feature case. The best recognition results are obtained in this experiment with a bandwidth θ_c around 8 Hz, a lower value than that resulting from the one feature case. Nevertheless, $v_1(n)$ has a narrower transition band than the $v_0(n)$ used in that case so that the effective bandwidth is similar in both cases.

L	13	14	15	16	17	18	19	20
Errors	17	13	14	13	14	12	13	15

Table 3. Recognition errors using three features and Legendre filters.

Table 3 shows the number of recognition errors using the first two Legendre filtered parameters (the conventional regression features) along with the unfiltered parameter, for various values of the length L. The results are worse in this experiment than those obtained with Slepian filters; however, Legendre filters only have a degree of freedom, the length L.

6. CONCLUSIONS

Frequency analysis and linear filtering were used in this paper to gain insights into the TSSPs and several processing techniques applied to them in the past. A filter design method was also proposed and it was applied to isolated word recognition by employing Slepian sequences. The recognition results, though preliminary, provide more evidence of the significance of the TSSP spectrum and its equalization.

ACKNOWLEDGEMENTS

The authors wish to thank R. Rose, F.K. Soong, C.H. Lee, and M. Rahim for their valuable suggestions and stimulating discussions throughout this work.

REFERENCES

[1] L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[2] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. ASSP*, Vol. 34, nº 1, pp. 52-59, Feb.1986.

[3] K.F. Lee, *Automatic Speech Recognition*, Kluwer Ac. Publ., 1989.

[4] T.H. Applebaum, B. Hanson, "Robust speaker-independent word recognition using spectral smoothing and temporal derivatives", *Proc. EUSIPCO'90*, pp. 1183-6, Sept. 1990.

[5] C-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini, A.E. Rosenberg, "Improved acoustic modelling for large vocabulary CSR", *Computer Speech and Language*, Vol.6, pp. 103-27, 1992.

[6] H. Hermansky, N. Morgan, A. Bayya, P. Kohn, "Compensation for the effect of the com. channel in auditory-like analysis of speech (RASTA-PLP)", *Proc. EUROSPEECH'91*, pp. 1367-70, Sept. 1991.

[7] K. Katagishi, H. Singer, K. Aikawa, S. Sagayama, "Feature extraction using a matrix coefficient filter for speech recognition", *Speech Communication*, Vol. 13, No. 3-4, pp.297-306, Dec. 1993.

[8] Y. Tohkura, "A weighted cepstral distance measure for speech recognition", *IEEE Trans. ASSP*, Vol. 35, No. 10, Oct. 1987.

[9] B. H. Juang, L.R. Rabiner, J.G. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Trans. ASSP*, Vol. 35, No. 7, pp. 947-53, July 1987.

[10] D. Slepian, "Prolate spheroidal wave functions, Fourier analysis, and uncertainty - V: The discrete case", *The Bell System Tech. Journal*, Vol. 57, No. 5, pp. 1371-1430, May-June 1978.