ON FREQUENCY AVERAGING FOR SPECTRAL ANALYSIS IN SPEECH RECOGNITION

Climent Nadeu
Fèlix Galindo
Jaume Padrell

Universitat Politècnica de Catalunya, Barcelona, Spain

ABSTRACT

Many speech recognition systems use logarithmic filter-bank energies or a linear transformation of them to represent the speech signal. Usually, each of those energies is routinely computed as a weighted average of the periodogram samples that lie in the corresponding frequency band. In this work, we attempt to gain an insight into the statistical properties of the frequency-averaged periodogram (FAP) from which those energies are samples. Thus, we have shown that the FAP is statistically and asymptotically equivalent to a multiwindow estimator that arises from the Thomson's optimization approach and uses orthogonal sinusoids as windows. The FAP and other multiwindow estimators are tested in a speech recognition application, observing the influence of several design factors. Particularly, a technique that is computationally simple like the FAP's one, and which is equivalent to use multiple cosine windows, appears as an alternative to be taken into consideration.

1. INTRODUCTION

To find a set of Q parameters that reliably represent the spectral envelope of a given speech frame, the widely used filter-bank-based speech parameterization techniques (e.g. mel-cepstrum [1]) usually estimate the filter-bank energies (FBE) through a weighted averaging of the periodogram samples (i.e. the samples of the square magnitude of the DFT of the windowed speech signal) that lie in each of Q frequency bands. These bands can be distributed along the frequency axis either uniformly or according to a non-linear frequency scale such as the mel scale.

If the set of weights used to compute the FBE is the same for each band, those band energies can be seen as samples (non-linear sampling if a mel scale is used) of a spectral estimate that results from convolving the periodogram with the weighting function. In the following, we will refer to that spectral estimate as the Frequency-Averaged Periodogram (FAP) (it actually is the so-called Daniell's periodogram [2]).

The FAP can be viewed as belonging to the family of multiwindow (MW) spectral estimators, i.e. those that result from averaging several periodograms, each one computed with a different window. Since Thomson's introductory work [3], various researchers have claimed good statistical properties for the MW estimates that are computed with a set of orthogonal windows, which result from Karhunen-Loève (KL) eigenequations. Some of their results have been discussed in [4], where it is experimentally shown that when not only both variance and frequency resolution of the estimator are considered but also time resolution is taken into account (note that the time resolution is actually involved in speech processing), the statistical performance of the FAP and that of the estimator arising from the MW-KL formalism are almost identical. For that reason, as well for the fact that the implementation of the FAP only requires the computation of one periodogram, and, furthermore, because it offers a great flexibility in defining the bands and the shape of the weights, the FAP appears as a quite practical choice for extracting the spectral parameters to be used in speech recognition.

In this paper, an attempt is firstly made to theoretically investigate the statistical properties of the FAP. Unfortunately, the FAP does not match the optimal MW-KL formalism. Therefore, we have studied the mathematical properties of the FAP by relating it with spectral estimators arising from that approach. From that, new MW alternatives for speech parameterization appeared. They were tested in a speech recognition application, observing the influence of several design factors.

2. MULTIWINDOW SPECTRAL ESTIMATION AND THE COMPOSITE SPECTRAL WINDOW

Since Thomson's work [3], several recent spectral analysis methods are based on the multiple window (MW) approach. Given a signal x(n) between n=0 and n=N-1, they estimate the power spectral density by averaging the periodograms that result from *K* orthonormal windows or tapers $v_k(n)$, $0 \le k \le K-1$, $0 \le n \le N-1$, which are optimal in a given way.

The set of Slepian windows or discrete prolate spheroidal sequences used in the Thomson's method can be described as arising from the Karhunen-Loève eigenequation that, written in the frequency domain, is [3]

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} Q(\theta) K_N(\omega-\theta) V_k(\theta)\, d\theta = \lambda_k\, V_k(\omega) \qquad (1)$$

where $0 \le k \le K-1$, $V_k(\omega)$ is the Fourier transform of $v_k(n)$, λ_k is the corresponding eigenvalue, and $K_N(\omega)$ is the Fourier transform of a rectangular window ranging from 0 to N-1.

The Slepian windows are obtained from the Karhunen-Loève eigenequation when the kernel Q(ω) of (1) is

$$Q(\omega)=\begin{cases}1, & -W<\omega<W\\ 0, & |\omega|>W\end{cases} \tag{2}$$

Approximately, the choice K=2NW allows to consider only the windows that have most of their energy inside the band [3]. Using shapes for Q(ω) others than the rectangular one, different families of windows follow from the integral equations (1). Those orthonormal windows are also orthogonal with respect to Q(ω) as weight. That orthogonality is searched in order to reduce the variance since orthogonal windows lead to uncorrelated periodograms.

In this paper, we are going to consider two non-rectangular the kernels: a triangular function and a function that is the square magnitude of the frequency response (spectral response) of a single-pole discrete system; to be concise, we will refer to the latter as the pole function

Every MW estimator computes an estimate $\hat{S}(\omega)$ of S(ω) by averaging in some way the power within a band surrounding the current frequency ω. Each window contributes to this average favoring some subbands in front of the others. If we wish to control that contribution, we have to assign different weights a_k to each windowed periodogram, i.e.

$$\hat{S}(\omega)=\sum_{k=0}^{K-1}a_k\left|\sum_{n=0}^{N-1}v_k(n)x(n)e^{-j\omega n}\right|^2 \tag{3}$$

so that $\sum_{k=0}^{K-1}a_k=1$.

In order to have a measure of the combined effect of the set of windows on the frequency domain, a *composite spectral window* (CSW) can be defined as [4]

$$W(\omega)=\sum_{k=0}^{K-1}a_k\left|V_k(\omega)\right|^2 \tag{5}$$

The CSW is a meaningful function for every MW estimator. In fact, as it is experimentally shown in [4], when the CSW of two MW techniques have a similar shape, the statistic performance of both estimators in terms of bias and variance is almost identical.

The weights a_k are arbitrary but if we choose them as normalized eigenvalues, i.e.

$$a_k=\frac{\lambda_k}{\sum_{j=0}^{K-1}\lambda_j} \tag{6}$$

it results that the functions Q(ω) and W(ω) of any given MW estimator show a similar shape. Specifically, it can be shown [7] that

$$W(\omega)\xrightarrow[N\to\infty]{}\frac{2\pi}{\int_{-\pi}^{+\pi}Q(\omega)d\omega}Q(\omega) \tag{7}$$

3. FREQUENCY AVERAGING AND SINUSOIDAL WINDOWS

The FAP technique computes a weighted average of K periodogram samples within a given band around the current frequency. As shown in the Mullis-Scharf's tutorial of quadratic estimators [5], it can be viewed as a MW technique, where the K windows result from multiplying a base rectangular window by complex exponentials that produce frequency shifts of its Fourier transform.

Some insight into the properties of the FAP can be obtained from the MW estimators that use sinusoidal windows. They arise from the KL eigenequations when Q(ω) is a pole function [6]. Particularly, as Riedel and Sidorenko have shown [7], the set of sine-wave windows has interesting properties since: 1) it is very close to the set of minimum bias windows, and 2) the computation of the spectral estimate can be implemented in a straightforward way since, given a frequency ω, the estimator is

$$\hat{S}_S(\omega)=\sum_{k=0}^{\frac{K-1}{2}}a_k\left|X(\omega-\omega_k)-X(\omega+\omega_k)\right|^2 \tag{4}$$

where X(ω) is the Fourier transform of x(n).

FAP uses complex exponential windows instead of sinusoids, which is equivalent to say that it applies expression (4) without the cross-term that is included in it. Figure 1 schematically shows the operations which define the FAP (average of the square magnitudes at each side of the central frequency), the sine case (square magnitude of the difference) and the cosine case (square magnitude of the addition).

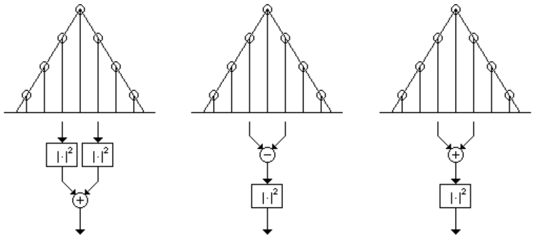

Figure 1. Operations involved in, from left to right: the FAP, the sine case and the cosine case.

Actually, the FAP estimate results from averaging the two MW estimates with sine and cosine windows. Therefore, it can be shown [6] that, asymptotically, i.e. when N tends to infinite, both the mean and the variance of the FAP estimator coincide with those of the MW estimators with sine and cosine. Moreover, the N exponential windows which are implicitly considered by the FAP technique are also orthogonal. In conclusion, we can say that, although the FAP does not result from the KL optimization framework, it is close to optimal estimators at least in a statistical and asymptotic sense.

The role of the CSW W(ω) of (5) is quite relevant for the properties of the MW spectral estimator [4] and it is determined by the weights a_k in (3). For estimators arising from the MW-KL formalism others than the sinusoidal ones, we will choose the weights according to (6), so the CSW will almost determine the identity of the windows set through its close relationship with the kernel Q(ω) (see expression (7)). Following the work by Hansson et al.[8], we also found in [4] that if we know a priori that the spectrum has prominent peaks, it will be useful in terms of bias and variance to employ a non-rectangular CSW. In fact, speech spectra show that kind of peaks so in speech processing it will be convenient to employ peaky shapes for W(ω), like the triangular shape that is conventionally used for computing the FBEs in speech recognition. Interestingly enough, we have observed [6] that the windows which arise from the MW-KL approach with a triangular Q(ω) have almost sinusoidal shapes.

4. APPLICATION TO SPEECH PARAMETERIZATION

Many speech recognition systems use logarithmic filter-bank energies or a linear transformation of them (e.g. mel-cepstrum) to represent the speech signal. Usually, each of those energies is routinely computed as a weighted average of the periodogram samples that lie in the corresponding frequency band so they can be seen as samples of the FAP. In the above sections we have shown the close relationship between FAP and estimators coming from an optimal MW approach. Therefore, it is not surprising that FAP is the most widely used spectral analysis technique for estimating the speech spectral envelope.

Anyway, our investigation has risen alternatives to FAP that will be tested in the next section. The most attractive ones are the MW estimators based on sinusoidal windows since they only require one DFT like FAP.

Several shapes will be considered for W(ω): rectangular, triangular and pole function. The last one may be the most appropriate if the aim of spectral estimation is to accentuate the spectral peaks since, according to the speech production model, formants have a shape like that of the pole function. Hence, it may be a good choice for formant estimation. However, we can not assert that enhancing spectral peaks is convenient for speech recognition, since they are not the formant frequencies that are used as spectral parameters but the samples of the spectral estimate at the centers of the various frequency bands. In fact, as we will see in the next section, the best results are not obtained with the pole function but with the triangular one.

In fact, a problem we face in speech recognition trying to improve just the spectral estimator in terms of statistical performance is that the posterior fixed sampling of the estimate does not take into account the actual spectrum of the current frame. Thus, an improvement in terms of –for instance– frequency resolution must not necessarily mean a better recognition performance.

5. SPEECH RECOGNITION RESULTS

We carried out several recognition tests with a telephone speech database of single Catalan digits. 2275 digit utterances were used for training and 2000 for testing. There was not any selection of files according to SNR, type of noise, dialect or speaker.

A speech recognition system based on continuous observation density hidden Markov models was used in the experiments (HTK 2.1). Each of the 11 digit models consisted of 8 (emitting) states, and the silence model had 3 states. For computational simplicity, only one diagonal covariance Gaussian pdf was employed per state.

Assuming as usual that, in a short-term basis, a stationary process can model the speech signal, a frame-to-frame spectral analysis yields a temporal sequence of spectral estimates that represent the acoustic-perceptual content. First of all, the 8 kHz sampled speech signal was Hamming windowed. Each frame was 32 ms long (256 samples) and the frame shift was 10 ms. After computing the spectral parameters for each frame of a given utterance, the average value of each time sequence of spectral parameters was removed from it (spectral mean removal). The delta parameters and the delta energy were also computed and included in all the tests.

Speech recognition tests were carried out for the various techniques presented in the paper, using the three different above mentioned CSW (rectangular, triangular and pole function), using either the uniform or the mel scale for the distribution of the bands between 100 Hz and 4000 Hz, and using two values for the number of bands Q: 12 and 24. Note that in the perceptively important frequency range between 100 and 1000 Hz, the uniform scale with 24 bands and the mel scale with 12 bands show a very close distribution of bands along the frequency axis.

In order to separate the effect of the spectral estimator from the subsequent linear transformation that is applied to the set of band energies, we carried out tests for two cases:

1) using just the filter-bank energies (FBE), without any posterior transformation;

2) applying to the FBE a frequency filtering operation (FBE-FF) with the usual filter $1\text{-}z^{-1}$[9].

First of all, speech recognition results for three spectral estimation techniques that use only one DFT will be presented. They correspond to the sets of exponential (FAP), sine (SIN) and cosine (COS) windows. The CSW are denoted with the capital letters R (rectangular), T (triangular), and P (pole function). UNIF and MEL refer to, respectively, the uniform and the mel frequency scales. The two numbers of bands for the uniform scale will be denoted by UNIF-12 and UNIF-24.

According to the results shown in Table 1, for the FBE case, both SIN and COS techniques outperform FAP’s one always, i.e. for every frequency scale and every CSW; however, when FF is used (Table 2), the difference is reduced. COS achieves the highest rates for both cases.

FBE		12 bands		24 bands
		UNIF	MEL	UNIF
FAP	R	84.74	86.89	83.89
	T	83.44	84.54	83.69
SIN	R	85.84	88.29	86.09
	T	85.34	88.74	84.24
	P	83.84	86.74	83.09
COS	R	85.29	89.49	85.44
	T	85.64	89.19	85.09
	P	82.34	87.94	83.99

Table 1: Recognition rates for FAP, SIN and COS without FF.

FBE-FF		12 bands		24 bands
		UNIF	MEL	UNIF
FAP	R	94.05	95.10	94.75
	T	94.05	96.10	96.20
SIN	R	95.10	95.00	95.20
	T	94.65	95.55	96.25
	P	93.90	95.25	96.40
COS	R	94.45	94.95	95.60
	T	94.05	96.70	96.90
	P	93.40	95.35	96.70

Table 2: Recognition rates for FAP, SIN and COS with FF.

Regarding the type of CSW, for FBE using any frequency scale and for FBE-FF using UNIF-12, R works, in general, equal or better than T, and T better than P. But for FBE-FF using either UNIF-24 or MEL, T and P perform always better than R. Note that the best results are obtained with the triangular window, the one that is conventionally used in speech recognition.

MEL performs much better than UNIF-24 bands for FBE, and FF reverses again the situation, since, in general, UNIF-24 gets slightly higher rates than MEL for FBE-FF. However, it requires double number of parameters.

Table 3 presents a few meaningful recognition tests carried out for the FBE-FF case, using MW techniques that can not avoid the use of several DFT computations and whose sets of windows arise from the KL formulation of Section 2 by choosing Q(ω)=W(ω). Only the uniform scale was employed since a non-uniform scale would require a different set of windows for each band. The number of windows taken is K=2NW for the rectangular CSW (as pointed out in Section 2), and K=4NW for the triangular one.

FBE-FF		UNIF	
		12 bands	24 bands
MW-KL	R	94.40	94.55
	T	93.85	96.20

Table 3: Recognition rates for the MW-KL estimator using FF.

Comparing results from Table 3 with those from Table 2, we see there is not much difference between FAP and MW-KL estimators that theoretically share the same CSW, but the results are not coincident. In fact, the actual shapes of their CSW are not exactly the same because they are fixed by different procedures. They are more similar for the triangular case, the one for which the recognition rates from both estimators are closer.

In conclusion, in our (preliminary) recognition experiments, the highest rates for all the techniques, using 12 spectral parameters per frame, were obtained with the mel scale and the triangular window. Notice that both are already employed in the conventional filter-bank speech parameterization. The MW-KL estimator performs similarly to FAP, confirming the decisive role of the CSW observed in [4]. COS obtains the highest scores so that it might be an alternative to FAP, but more experiments must be carried out to test it in diverse situations.

6. CONCLUSIONS

In this paper, an attempt has been made to theoretically investigate the statistical properties of the FAP and we have shown that, asymptotically and in terms of the first and the second moments of the estimator, the FAP is equivalent to the MW-KL estimator that uses orthogonal sinusoids as windows. This fact suggested us that the MW-KL spectral estimator that is based on sinusoidal windows could also be used for speech recognition. Actually, it does not require more computations that the FAP since it can also be obtained from only one DFT, and it also performs a kind of frequency averaging, but using additional cross products of periodogram samples. Speech recognition tests were carried out for these and the other MW techniques considered in the paper. The results show that the cosine technique appears as an alternative to be taken into consideration.

7. ACKNOWLEDGMENTS

This work has been funded by the Spanish government projects TIC95-1022-C05-03 and TIC95-0884-C04-02. The digit database has been used with permission from *Telefónica Investigación y Desarrollo* and *Fundació Catalana per a la Recerca*.

8. REFERENCES

1. S.B. Davis, P. Mermelstein, *IEEE Trans. on ASSP*, Vol. ASSP-28, No.4, pp. 357-366, August 1980.
2. S.L. Marple, *Digital Spectral Analysis with Applications*, Prentice-Hall, 1987.
3. D. Thomson, *Proc. IEEE*, Vol. 70, No. 9, pp. 1055-96, Sept.1982.
4. C. Nadeu, J. Padrell, I. Esquerra, *Proc. ICASSP'97*, pp. 3953-6, April 1997.
5. C.T. Mullis, L.L.Scharf,, *in Adv. in Spectrum Analysis and Array Processing*, S. Haykin, Ed., Prentice-Hall, 1990.
6. F. Galindo, *Final Project for the Telecommunication Engineering Diploma*, 1998.
7. K. Riedel, A. Sidorenko, *IEEE Trans. on SP.*, Vol. 44, No. 7, July 1996.
8. M. Hansson, T. Gänsler, Göran Salomonsson, *Proc. ICASSP'95*, Detroit, pp. 1617-20, May 1995.
9. C. Nadeu, J. Hernando, M. Gorricho, *Proc. EUROSPEECH'95*, pp. 1381-4, Sept. 1995.