Improving the robustness of the usual FBE-based ASR front-end

Climent Nadeu, Dušan Macho, and Javier Hernando

TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain
Email: {climent,dusan,javier}@talp.upc.es

Abstract

All speech recognition systems require some form of signal representation that parametrically models the temporal evolution of the spectral envelope. Current parameterizations involve, either explicitly or implicitly, a set of energies from frequency bands which are often distributed in a mel scale. The computation of those filter-bank energies (FBE) always includes smoothing of basic spectral measurements and non-linear amplitude compression. A variety of linear transformations are typically applied to this time-frequency representation prior to the Hidden Markov Model (HMM) pattern-matching stage of recognition. In the paper, we will discuss some robustness issues involved in both the computation of the FBEs and the posterior linear transformations, presenting alternative techniques that can improve robustness in additive noise conditions. In particular, the root non-linearity, a voicing-dependent FBE computation technique and a time&frequency filtering (tiffing) technique will be considered. Recognition results for the Aurora database will be shown to illustrate the potential application of these alternatives techniques for enhancing the robustness of speech recognition systems.

1. Introduction

Current speech recognition systems use a pattern matching approach [35]. The classifier, which is commonly based on hidden Markov models (HMM), relies on a speech spectrum representation that must be robust to signal degradations.

The most widely used spectral parameters are the logarithmic filter-bank energies (log FBEs). Usually, the discrete cosine transform (DCT) is used to compute from the log FBEs a set of uncorrelated parameters, the so-called mel-frequency cepstral coefficients (MFCC) or mel-cepstrum, probably the most used spectral representation in speech recognition [3]. On the other hand, orthogonal (Legendre) polynomial filters are used to compute the supplementary dynamic (delta) feature vectors for each frame [5]. For example, the recent distributed speech recognition (DSR) standard front-end for clean speech (ETSI STQ WI007 [4]) establishes this kind of speech representation.

In this paper, we intend to address several issues involved in the various blocks of the parameterization scheme, reviewing and discussing a few salient points that have traditionally been considered unquestionable. In particular, both the logarithmic non-linearity, the cepstral coefficients and the usual delta and double-delta time filters will be discussed and more robust alternatives will be presented.

2. Non-linearly compressed filter-bank energies

In current speech recognition front-ends, the first parameterization step consists of extracting a short-time representation of the spectral envelope for each speech frame. There are many reported techniques to estimate the set of spectral parameters [15][34][35], but they always combine some kind of smoothing of raw spectral measurements with non-linear operations.

2.1 Spectral smoothing and FBEs

Spectral smoothing is used to remove the harmonic structure of the speech spectrum corresponding to pitch information and to reduce the variance (error) of the speech spectral envelope estimation. Additionally, an envelope representation with a small number of parameters is obtained. That operation has basically been done in two alternative ways: linear prediction (LP) analysis and spectral band energy estimation [35]. The strength of the LP method arises from the fact that it matches the all-pole model of speech production. In this way, it is able to approximately separate the vocal tract response, which corresponds to the spectral envelope, from the glottal excitation.

However, the band energy parameters have become increasingly popular. They separately represent the energy at each frequency band since they result from integrating the energy values in the time-frequency area specified by the frame length and the effective bandwidth. The main reason of the usefulness of these energies is perhaps the higher flexibility of the sub-band approach with respect to the full-band approach involved in LP modeling. In fact, it offers the possibility of defining the width and shape of the bands along the frequency axis. Also, if the signal-to-noise ratio (SNR) of each band is known, the band energy representation allows to use it in straightforward ways: noise masking, spectral subtraction, etc [15].

Currently, the most used implementation of the filter-bank analysis [3] operates in the frequency domain by computing a weighted average of the magnitude (or, sometimes, the squared magnitude) of the DFT values of the windowed speech frame in each frequency band, obtaining in this way the so-called filter-bank energies (FBEs). Figure 1 shows the sequence of operations involved in the computation of the FBEs for a given windowed speech frame; it also includes the posterior non-linear compression step from section 2.3.

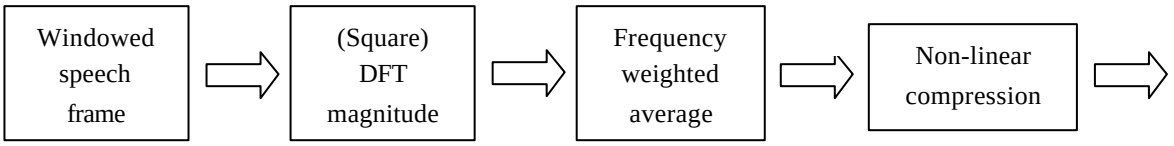

Figure 1. *Usual scheme for computing the non-linearly compressed filter-bank energies for a given frame. Sometimes, a LP modeling block is inserted at the end (like in PLP [8]).*

LP speech spectral estimates are well established theoretically, since they are based on the all-pole model of speech production. However, the theoretical foundations of the above mentioned FBEs (from [3]) have not received much attention so far, in spite of they have become a kind of standard in speech recognition. In [31] the authors show that, assuming an uniform frequency scale, the FBEs come from a spectral estimator that matches the multiwindow formulation introduced by Thomson in [37] and shows good statistical properties. In fact, it is equivalent, asymptotically and in terms of the first and the second moments of the estimator, to the optimal multiwindow estimator that uses orthogonal sinusoids as windows [25].

2.2 Use of voicing information to improve the robustness of the spectral parameter set

In general, the effect of additive noise on the speech spectrum is more remarkable at frequency bands or time segments where the speech spectrum shows low amplitude, e.g. the between-harmonic valleys of voiced sounds and the whole band of unvoiced (or silence) sounds.

If the spectrum is expressed in dB, the added noise relatively increases the spectral values at those low-power spectral regions or time segments more than it does with the high-power ones, so the amplitude contrast of the spectrum decreases along frequency and also along its time evolution. However, the voicing information can be used to restore that amplitude contrast up to some extent. This was the approach taken in [22], where the character of each frame is introduced explicitly in the computation of the FBEs by means of an exponent that depends on it.

Tests have been carried out for two training modes: clean speech training, and multicondition training (i.e. the training corpus contains both clean and noisy speech signals, for various noise conditions). The Aurora 1.0 database, that is being used for developing the ETSI noisy speech Distributed Speech Recognition (DSR) standard front-end [33] was employed in this work for testing the described exponentiation techniques. This corpus consists of the TI connected digit utterances, downsampled from 20 kHz to 8 kHz, and with artificially added noises at the following SNR levels: clean, 20, 15, 10, 5, 0, and –5 dB. Four types of real noises have been used: hall, babble, suburban train, and car. Noise conditions 0 and –5 dB are not used for training, and an average word recognition accuracy for noisy speech is computed by considering all the test conditions for noisy speech except –5 dB. The recognition system, which is the one used for the above standardization work, is based on continuous density HMMs with diagonal covariance matrices.

Significant recognition improvement was obtained when using static parameters alone (without time derivatives) by employing any of two different exponentiation techniques. By comparing both techniques, it was concluded that the predominant effect is the enhancement of the amplitude contrast between voiced and unvoiced/silence segments in the temporal sequences of spectral parameters.

2.3 Non-linear compression

To compute the speech parameters, non-linear processing is used in both axis of the spectral representation. First, the bands are often distributed in a mel scale to mimic the properties of the human auditory processing, giving less emphasis to the high-frequency bands. And, secondly, non-linear operators are used to compress the large amplitude range of spectral measurements, producing a distribution more similar to the Gaussian one.

The most used non-linear operator is the logarithm which has the additional advantage of converting a gain factor in an additive component in the feature space, which can be easily removed. Although the logarithm is perhaps the most appropriate non-linear operator for recognition of clean speech, it may no longer keep its advantage whenever additive noise is present. Other reported non-linear operators, such as the root $|E|^g$ [1] or the lin-log *log(1+JE)* [10], where *E* denotes a spectral measurement (usually a FBE), are alternative candidates to cope with the problem of parameterizing noisy speech. Actually, both have a parameter which can be adapted to the SNR: g [38] or J [10]. Recently, both techniques were interpreted as masking procedures at spectral valleys [14].

In the following, some aspects of the non-linearly compressed spectrum of speech when it is corrupted by additive white noise will be discussed. Assuming that the speech signal and the noise are uncorrelated, the spectrum $E_N(w)$ of a distorted speech frame is the sum of both speech spectrum $E(w)$ and noise spectrum $N(w)$

$$E_N(w) = E(w) + N(w) \tag{1}$$

This additive property is not longer true when a non-linearity is applied to the speech spectrum. When the logarithm is applied to the noisy speech spectrum, it follows that

$$\log(E_N(w)) = \log(E(w) + N(w)) = \log(E(w)) + [\log(E(w) + N(w)) - \log(E(w))] = \log(E(w)) + \log\left(1 + \frac{N(w)}{E(w)}\right) \tag{2}$$

The new noise term $\log(1 + N(w)/E(w))$ becomes speech dependent so its statistics vary in frequency and time according to the speech spectrum variations. We studied the noise term distribution in a logarithmically compressed spectrum of a stationary segment of speech that was assumed to be deterministic (by adding different frames of white noise to the center frame of a phoneme) [24].

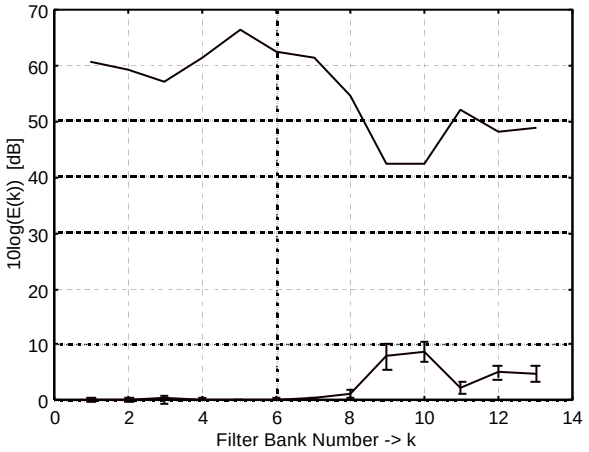

Figure 2 *13 log FBE spectral representation of clean phoneme /e/ (the upper curve) and the mean and variance distribution of its noisy term log(1+N(k)/E(k)) (the lower curve). E(k), N(k) are the k-th band energies of speech and noise, respectively.*

Figure 2 shows the distribution of both the mean and the variance of the noise term in the log FBE spectral estimation. The spectral valleys are more contaminated than the spectral peaks. Let us show that this is a general fact by differentiating the noise term in (2)

$$\frac{d}{dw} log\left(1+\frac{N(w)}{E(w)}\right)=\left(-\frac{N(w)}{(E(w)+N(w))E(w)}\right)\frac{dE(w)}{dw} \quad (3)$$

The term in parenthesis, which is multiplied by the clean speech spectrum derivative, is always negative for $E(w), N(w) > 0$. This implies that the derivative of the noise term has always an inverse sign to that of the derivative of the clean speech spectrum. Therefore, where there is a local maximum/minimum in the clean speech spectrum, the noise term has a local minimum/maximum, respectively, so spectral valleys are more affected by noise than peaks. The same conclusion can be derived in the case of the root non-linearity.

Consequently, by using a non-linearity that enhances peaks with respect to valleys more than the logarithm does, more robust parameters may be obtained. With the root non-linearity, the degree of compression of the input dynamic range can be controled by the parameter g. In [23], the behaviour of the root and log non-linearity functions was investigated (along with that of two additional non-linearities: generalized logarithm and lin-log) by using a quotient D of their derivatives at two points of the variable (the spectral value): the first larger than the second. A higher quotient D means that for larger spectral values the slope of the non-linearity is higher than for smaller spectral values. The results are:

1. Those quotients are independent on the considered dynamic range of spectral values
2. When $g \to 1$ then $D_{Root} \to 1$, and if $g \to 0$ then $D_{Root} \to D_{Log}$
3. For all $g \in (0,1)$, $D_{Root} > D_{Log}$

Based on our previous discussion and observing the last result, we can conclude that the use of the root non-linearity can yield more robust parameters (and the same conclusion can be reached for the generalized log and the lin-log). On the other hand, the discrimination capability must be also considered. Actually, a g too high, which seems to be a good robust solution, can cause the distribution of parameters is not longer Gaussian-like and the recognition performance can decrease. Therefore, there exists a trade-off in the g selection and g should be set up empirically by recognition tests.

A trade-off is usually observed when comparing the differences in recognition performance for clean and noisy speech either between two different values of the control parameter or between two non-linearities. For example, a higher γ value is more suited to speech contaminated with white noise, but it is less suited to clean speech [24]. And the root can achieve better recognition results than the logarithm for white or broad-band noises, whereas the log performs better than the root for clean speech or speech-like noise [23].

3. Linear transformations of the spectral energies

The vector of Q log FBEs is linearly transformed in order that the feature vectors supplied to the pattern matching stage are better adapted to the assumptions of the HMM formalism and take more advantage from it. That vector undergoes at least two kinds of linear transformations, one in the frequency domain and the other in the time domain. In the next sections, both kinds of transformations will be discussed, and also alternatives for improving the recognition performance will be proposed.

3.1 The frequency filtering (FF) technique

Usual HMMs assume that the acoustic observation vectors can be modeled by Gaussian distributions with diagonal covariance matrices, i.e. they assume that the elements of those vectors are uncorrelated. As the spectral measurements are strongly correlated (e.g. the correlation coefficient of log FBEs of adjacent bands for the TI digits database [17] and Q=12 is 0.92 [29]), the parameterization front-ends require a linear transformation that obtains a set of spectral parameters that are globally decorrelated. Usually, a discrete cosine transform (DCT) is employed for that purpose. Due to its closeness to the optimal K-L transform, the DCT is able not only to nearly decorrelate the vector of logarithmically compressed FBEs but also to sort the transformed coefficients in variance order. Then, the resulting vector is truncated to retain the highest energy coefficients. It is the mel-frequency cepstral coefficients (MFCC) representation, also called mel-cepstrum. That truncation actually represents an implicit liftering operation with a rectangular lifter that smoothes the spectral envelope represented by the frequency sequence of Q log FBEs $S(k)$.

The cepstral coefficients show two important disadvantages for speech recognition:

1. They do not lie in the frequency domain, so lacking a frequency meaning which may be useful, especially for implementing robust techniques.

2. As most current HMMs use Gaussian distributions with diagonal covariance matrices and ML-estimated standard deviations, those HMMs can not benefit from a cepstral weighting (liftering), since any multiplying factor that is applied to the observations does not affect the Gaussian exponent calculation.

An alternative set of speech parameters that avoids these disadvantages has been recently presented by the authors [26]. This new parameter set is obtained with a very simple linear transformation, called frequency filtering. Frequency filtering is a transformation of the parameter vector that consists of a convolution between the sequence *S(k)* of *Q* band log energies and a given (impulse response) sequence *h(k)* to obtain a new sequence of *Q* filtered parameters *F(k), k=1,...,Q,* i.e.

$$F(k) = S(k) * h(k),\ k = 1,\ldots,Q \tag{4}$$

Notice that the filtered parameters *F(k)* still lie in the frequency domain, and only *Q* values are computed. The sequence *h(k)* typically has order one or two, so the computational burden is minimal. It is designed to fulfill a double requirement [26]: 1) to decorrelate the parameters, and 2) to enhance their discriminative ability. As shown in [26], decorrelation can be approximately obtained with a derivative-type filter of order one which flattens the variance of the cepstral coefficients (the Fourier counterpart of the log FBEs), since filtering in frequency is equivalent to weighting the cepstral coefficients with the DFT of *h(k),* and the cepstral variance decreases along its index according to an one-pole spectral shape.

It is known that this kind of weighting, that is referred to liftering, can improve discrimination provided that it shows an increasing curve at least up to the 6th or 8^{th} cepstral coefficient. Actually, cepstral coefficients with middle indexes should be emphasized since they correspond to the "formant rate", i.e. the number of formants per period in the frequency axis. Thus, the same filter that approximately decorrelates the frequency sequence shows a liftering shape that can enhance discrimination.

Thus, the impulse response *h(k)* may be designed to maximally flatten the variance of the cepstral coefficients, or alternatively, its 1 or 2 coefficients may be empirically tuned to obtain the best recognition results. However, the simple data-independent second order filter $H(z)=z-z^{-1}$ (i.e. $h(k)=\{1,0,-1\}$) has shown a rather good performance for a wide range of conditions. The two endpoints of the filtered sequence actually are absolute energies, not differences, so the full-band energy may be neglected. In the FF approach, the number of bands is the number of transformed parameters as well, so it has to be carefully chosen.

It is worthing to note that the outputs *F(k), k=2,...,Q-1* of such a derivative-type filter actually are spectral slope measures and, according to Klatt, a phonetic distance based on the spectral slope near the peaks correlates very well with perceptual data, unlike other speech characteristics such as the FBE values or the linear prediction residual [16].

Many recognition experiments have been performed in our laboratory during the last years to assess the FF technique: for different speech recognition tasks (digit recognition and acoustic-phonetic decoding [26][28], word spotting with phone units [29]), for different noise conditions [12], for speaker recognition [11], and also using features that were not obtained from an usual filter-bank but from LP modeling [26][12]. From the whole set of tests, it appears that FF generally offers better recognition performance than MFCC.

Summarizing, we can conclude that frequency filtering is a simple and effective operation that performs a combination of decorrelation and liftering, while still maintaining the speech parameters in the frequency domain, so avoiding the above mentioned drawbacks of cepstral coefficients. Note in particular that FF coefficients may be especially useful whenever their frequency localization property is convenient. For instance, to use them in a missing feature paradigm, like in [39].

3.2 Robust temporal filtering and the modulation spectrum

The pattern-matching formalism based on HMM assumes that each acoustic observation vector is uncorrelated with its temporal neighbors. This assumption can not be fulfilled by the transformed vectors for the usual frame shifts (typically, 10 ms). That has been the reason to justify the inclusion of smoothed time derivatives as additional parameter vectors (they are also referred to as "dynamic" features [5]). Thus, not only the first-order differential parameter vector but often the second-order one are appended to the basic "static" vector (for the sake of simplicity of the explanation, we will assume in the following that the global energy, if used, and its differences are already included in the parameter vectors). These two new temporal sequences of differential vectors are computed by filtering the basic time sequence of spectral parameter vectors.

Filtering of each time sequence of spectral parameters (TSSP) has also been used for robust speech recognition with another goal: to remove its d.c. and slowly variant components when they are carrying undesired

perturbations as linear distortion (convolutional noise, additive in the log spectral domain). That is the aim of the cepstral mean subtraction (CMS) technique [36] and it is what basically does the IIR filter with a pole close to 1 that is used in the so-called RASTA processing [10].

The effect of temporal filtering (TF) can be better understood in the frequency domain. The frequency counterpart of the frame index n is the modulation frequency q [13]. For this reason, the TSSP spectrum has been called modulation spectrum (MS) [6]. In [30], from the analysis of the MS of filtered TSSP of clean speech, it was concluded that:

1. Each dynamic TSSP emphasizes a given band of meaningful modulation frequencies. This effect is achieved with an approximate equalization of the static MS in that band.
2. The modulation frequency bands of the various TSSP (static and dynamic) are distributed along an interval of the modulation frequency axis in such a way that the function that results from adding their MS is rather flat in that interval, which is phonetically relevant and does not carry an excessive spectral estimation noise.

It is not a fact under discussion that time-filtered features are less affected by convolutional noises than the static features, so that they can help the recognition system to cope with mismatches between training and testing data. However, that is not so clear for additive noises [7]. In [28], some recognition results were presented which give further evidence that the dynamic features can be less affected than static features by additive noises, provided that the time filters are properly designed, and that the modulation spectrum gives useful insight for that filter design.

In the above mentioned work, filter design has been based on an experimental approach; statistically optimal designs can be alternatively pursued, either based on a linear discriminant analysis approach [2][9] or on a maximum likelihood approach [32].

3.3 *Tiffing* (*TI*me and *F*requency *F*il*tering*)

Let us consider the two-dimensional (2-D) sequence of log FBEs $S(k,n)$, where the index k denotes the frequency band and the index n denotes the time frame. In the above sections, we have presented filtering as being separately performed in the dimensions k and n. However, the effect of filtering is remarkedly similar in both dimensions. In fact, time and frequency filters show similar characteristics since both perform a kind of smoothed derivative. Concretely, both the frequency filter and the time filters used in this work can be viewed as the combination of three operations: 1) removal (or severe attenuation) of the average value; 2) approximate variance or power equalization in the transform domain (quefrency for k, or modulation frequency for n) with a first-order high-pass FIR filter; and 3) smoothing of the resulting sequence with a low-pass filter that shapes the (equalized) band. Additionally, the effects of both kind of filters are not orthogonal; for example, the d.c. component of the 2-D time-frequency sequence $S(k,n)$ may be removed by both filters.

On the other hand, frequency-filtered log FBEs seem more able to benefit from temporal filtering than cepstral coefficients [28]. These observations lead us to think that there is something of a synergy effect between both types of filtering operations. Consequently, both types of filters can be considered together as applied to a two-dimensional frequency-time sequence, and the 2-D modulation spectrum (2D-MS) [19], can be helpful for designing and analyzing them. Several recognition tests were performed for the TI single digits database and 10 dB white noise. The best results were achieved using a different frequency filter for each time filter.

To conclude this section, let us mention a recent comparison between tiffing and the usual MFCC parameterization (with delta time filters) in the framework or the Aurora task mentioned in Section 2.2. The FF technique was used with Q=13 bands, so both parameterizations use the same number of parameters (MFCC includes 12 cepstral coefficients plus energy). The relative improvements of average accuracy rates were found meaningful: 25,8% and 9,8% for clean and noisy recognition, respectively, in comparison to the standard MFCC front-end. Furthermore, tiffing outperformed in average terms the usual mel-cepstrum representation for every kind of noise and SNR.

4. Acknowledgements

The authors would like to thank to E. Batlle, F. Galindo, J. Marí, J.B. Mariño, P. Pachès and J. Padrell for their helpful comments and discussions. This work has been supported by CICYT, projects TIC98-0683 and TIC98-0423-C06-01.

5. References

[1] Alexandre P., Lockwood P., "Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments", Speech Communication, Vol. 12, No. 3, 1993, pp. 277-288.

[2] Avendaño C., van Vuuren S., Hermansky H., "Data based filter design for RASTA-like channel normalization in ASR", Proc. ICSLP, 1996, pp. 2087-2090.

[3] Davis S.B., Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on Acoustic, Speech and Signal Processing, Vol. ASSP-28, No.4, pp. 357-366, August 1980.

[4] ETSI SQL W1007, http://webapp.etsi.org/WorkProgram/Report_WorkItem.asp?WKI_ID=6400

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

[6] Greenberg, S., Kingsbury, B.E.D., "The Modulation Spectrogram: in Pursuit of an Invariant Representation of Speech", Proc. ICASSP, 1997, Vol. 3, pp. 1647-50.

[7] Hanson B.A., Applebaum T.H., Junqua J.C., "Spectral Dynamics for Speech Recognition under Adverse Conditions", in: Lee, C.H., Soong, F.K., Advanced Topics in Automatic Speech and Speaker Recognition, Kluwer Academic Publisher, Dordrecht, 1996.

[8] Hermansky H., "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Amer. 87 (4), 1738-1752, 1990.

[9] Hermansky, H., "Should Recognizers Have Ears?", Speech Communication, No. 25, 1998, pp 3-27.

[10] Hermansky, H., Morgan, N., "RASTA Processing of Speech", IEEE Trans. on SAP, Vol. 2, No. 4, 1994, pp. 1-12.

[11] Hernando, J., Nadeu C., "CDHMM Speaker Recognition by Means of Frequency Filtering of Filter-Bank Energies", Proc. Eurospeech, 1997, Vol.5, pp.2363-2366.

[12] Hernando, J., Nadeu, C., "Robust Speech Parameters Located in the Frequency Domain", Proc. Eurospeech, 1997, Vol. 1, pp. 417-20.

[13] Houtgast, T., Steeneken, H.J.M., "A Review of the MTF Concept in Room Acoustics and its Use for Estimating Speech Intelligibility in Auditoria", Journal of Acoustic Soc. of America, No. 3, Vol. 77, 1985, pp. 1069-77.

[14] Hunt, M.J., "Spectral Signal Processing for ASR", Proc. Workshop ASRU, 1999.

[15] Junqua J.-C., Haton J-P., *Robustness in automatic speech recognition*, Kluwer, 1996.

[16] Klatt, D.H. "Prediction of perceived phonetic distance from critical band spectra: A first step", Proc. ICASSP, 1982, pp.1278-81.

[17] Leonard, R.G., "A Database for Speaker-Indepemdent Digit Recognition", Proc. ICASSP, Vol. 3, 1984, pp. 42-45.

[18] Lockwood, P., Alexandre, P., "Root Adaptive Homomorphic Deconvolution Schemes for Speech Recognition in Noise", Proc. ICASSP, Vol. 1, 1994, pp. 441-444.

[19] Macho, D., Nadeu, C., "On the Interaction between Time and Frequency Filtering of Speech Parameters for Robust Speech Recognition", Proc. ICSLP, 1998, pp. 1487-90.

[20] Macho, D., Nadeu, C., Hernando, J., Padrell, J., "Time and Frequency Filtering for Speech Recognition in Real Noise Conditions", Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere 1999, pp. 111-114.

[21] Macho, D., Nadeu, C., Jancovic, P., Rozinaj, G., Hernando, J., "Comparison of Time & Frequency Filtering and Cepstral-Time Matrix Approaches in ASR", Proc. Eurospeech 1999, Vol. 1, pp. 77-80.

[22] Macho, D., Nadeu, "Use of voicing information to improve the robustness of the spectral parameter set", Proc. ICSLP, 2000.

[23] Macho, D., PhD Thesis Dissertation, 2000.

[24] Marí, J., Engineering Degree Project, ETSETB, Universitat Politècnica de Catalunya, Barcelona, 1997.

[25] Nadeu, C., Galindo, F., Padrell, J., "On Frequency Averaging for Spectral Analysis in Speech Recognition", Proc. ICSLP, 1998, Vol. 3, pp 1071-74.

[26] Nadeu, C., Hernando, J., Gorricho, M., "On the Decorrelation of Filter-Bank Energies in Speech Recognition", Proc. Euro speech, 1995, pp. 1381-84.

[27] Nadeu, C., Juang, B.H., "Filtering of Spectral Parameters for Speech Recognition", Proc. ICSLP, 1994, pp. 1927-30.

[28] Nadeu, C., Macho, D., Hernando, J.., "Time&Frequency filtering of FBEs for Robust HMM Speech Recognition", to appear in Speech Communication, 2000.

[29] Nadeu, C., Mariño, J.B., Hernando, J., Nogueiras, A., "Frequency and Time Filtering of Filter-Bank Energies for HMM Speech Recognition", Proc. ICSLP, 1996, pp. 430-433.

[30] Nadeu, C., Pachès-Leal, P., Juang, B.H., "Filtering the Time Sequence of Spectral Parameters for Speech Recognition", Speech Communication, Vol. 22, 1997, pp. 315-322.

[31] Nadeu, C., Padrell, J., Esquerra, I., "Frequency Averaging: An Useful Multiwindow Spectral Analysis Approach", Proc. ICASSP, 1997, pp. 3953-56.

[32] Pachès-Leal P., Rose R.C., Climent Nadeu, C., "Optimization Algorithms for Estimating Modulation Spectrum Domain Filters", Proc. Eurospeech, 1999, Vol. 1, pp. 89-92.

[33] Pearce, D., "Experimental Framework for the Performance Evaluation of Distributed Speech Recognition Front-Ends", Aurora project report, Version 1, Sept. 1998.

[34] Picone, J.W., "Signal Modeling Techniques in Speech Recognition", Proc. of the IEEE, Vol. 79, No. 4, 1991, pp. 1214-1247.

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

[36] Rosenberg, A.E., Lee, C.-H., Soong, F., "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", Proc. ICSLP, 1994, 1835-9.

[37] Thomson, D.J., "Spectrum Estimation and Harmonic Analysis", Proc. of the IEEE, Vol. 70, No. 9, 1982, pp. 1055-1096.

[38] Tian, J., Viikki, O., "Generalized Cepstral Analysis for Speech Recognition in Noise", Proc. Eurospeech, 1999, pp. 87-90.

[39] Veth J. de, Wet F. de, Cranen B., Boves L., "Missing Features Theory in ASR: Make Sure You Miss the Right Type of Features", Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, 1999, pp. 231-4.

[40] Young, S., Jansen, J., Odell, J., Ollason, D., Woodland, P., The HTK Toolkit.