A COMPARATIVE STUDY OF SEVERAL PARAMETRIC REPRESENTATIONS FOR SPEECH RECOGNITION[1]

L. J. Rodríguez[2] and M. I. Torres

Departamento de Electricidad y Electrónica. Universidad del País Vasco.
Apartado 644. 48080 BILBAO. SPAIN.
e-mail : luisja@we.lc.ehu.es, manes@we.lc.ehu.es

Abstract-. Discriminative acoustic features of speech signal are represented through different sets of spectral parameters in speech recognition tasks. Several parametric representations were considered in this work: log-area ratios, reflection and cepstral coefficients obtained through Linear Prediction analysis, and Fourier Transform derived cepstral coefficients. All of them were evaluated over an easy recognition task. Their statistical behaviour as well as the influence over the system performance of vector size, bilinear transformation coefficient, pre-emphasis and energy, was also evaluated along with the analysis method.

Introduction-.

Automatic Recognition of Continuous Speech is nowadays a classical scope of Pattern Recognition. From this perspective, speech can be considered as composed by a relatively small set of primitives. Some combination of these primitives in a multilevel hierarchy would lead to a large set of more complex patterns which are strongly related by a powerful structure [Levinson, 85]. The spectral analysis of the speech signal constitutes the zero level of the above hierarchy. The mapping between the acoustic features and further sublexical units (phones, syllables, context-dependent phones, etc.) can then been automatically learnt from a training set, typically within the Syntactic Pattern Recognition approach.

The object of the spectral analysis procedure is to provide relevant information that should be speaker, task and environment-independent. The discriminative acoustic features of speech signal are finally represented by a small set of robust parameters. The evaluation is usually based on final recognition rates. However, the statistical behaviour should also be considered when choosing among several possible parametric representations. Some approaches, like semicontinuous and continuous hidden Markov modelling, assume the parameters to be statistically independent, thus having diagonal covariance matrices [Bellegarda, 90][Huang, 93]. On the other hand, distance-based methods assume a similar variance for all the coefficients [Nocerino, 85][Tohkura, 87]. These assumptions are more or less adequate depending on the analysis procedure.

The aim of this work was to evaluate a large set of parametric representations over an easy recognition task. An analysis of their statistical behaviour was also performed. Finally

[1]Work partially supported by the Spanish CICYT under grant TIC94-0210-E.

[2]Supported by the *Gobierno Vasco* under grant BFI93.092-AE.

we analyzed the influence of several factors in the recognition performances: analysis method, bilinear transformation, pre-emphasis, energy and number of coefficients.

Parameters-.

Two groups of parametric representations were compared (Table I): Linear Prediction (LP) and Fourier Transform (FT). The first one is considered as the best choice for 8 and 10 kHz sampled databases [Lee, 89], and includes reflection coefficients (RC), log-area ratios (LAR) and cepstral coefficients (LPCEP). These parameters are equivalent representations of the LP spectrum. Typical vector sizes are 14 for RC and LAR, and 12 for LPCEP [Lee, 89].

The second group includes the Fast Fourier Transform derived cepstral coefficients (FFTCEP), as defined in [Rabiner, 78]. The perceptual frequency distortion represented through the Bark scale [Zwicker, 81] can be introduced by averaging the FFT coefficients in consecutive bands of frequency (BFB). The number of bands is related to the sampling frequency: 21 bandpass filters are needed for a 16 kHz sampling rate. Then the corresponding cepstral coefficients (BFBCEP) can be computed by applying a cosine transform.

The BFB derived parameters have been typically used for speech recognition with sampling rates of 16 kHz, whereas LP ones have been used for 8 and 10 kHz [Shikano, 86][Lee, 89]. However, there is no experimental evidence for this choice. In this work we attempted to solve this point by a direct comparison of system performances in an isolated word recognition task. The Bilinear Transformation (BLT) of the time sequence [Oppenheim, 72][Lee, 89] was introduced for FFTCEP and LPCEP to emulate the perceptual frequency distortion represented by the Bark scale.

Another problem which has not been completely solved [Paliwal, 84] is the role of preemphasis in speech analysis for recognition. Therefore, all the parameters were computed with and without preemphasis of the speech signal. A Hamming window of 32 ms, with overlapping of 16 ms, was applied in all cases. Table I shows a list of the parameter sets considered in this work.

Table I. Parametric representations tested: LP-based (I) and FT-based (II). Preemphasis was considered in all cases. Bilinear transformation (BLT) was applied only to LPCEP and FFTCEP.

PARAMETER	vector size	preemphasis	BLT	log-energy
I.1-. Reflection Coefficients (RC)	14	Y/N	NO	NO
I.2-. Log-Area Ratios (LAR)	14	Y/N	NO	NO
I.3-. LP derived cepstral coefficients (LPCEP)	12	Y/N	Y/N	Y/N
II.1-. FFT derived cepstral coefficients (FFTCEP)	12	Y/N	Y/N	Y/N
II.2-. Bark-scaled Filter Bank coefficients (BFB)	21	Y/N	NO	NO
II.3-. BFB derived cepstral coefficients (BFBCEP)	12	Y/N	NO	Y/N

The log-energy of each frame was computed, re-scaled and alternatively added as the first component to the vector of parameters in the following cases: BFBCEP, LPCEP and FFTCEP, with different vector sizes: 6, 8, 10 and 12.

Experiments and Results-.

An experimental evaluation was carried out on an easy recognition task. The corpus consisted of 1000 utterances of the Spanish digits uttered by ten speakers and was acquired at 16 Khz. In order to obtain a significant size for the test set, a 5-fold cross-validation technique was applied [Raudys, 91]. In each partition the training size was 800 utterances and the test set was 200 utterances, yielding an effective test set of 1000 utterances. The recognition procedure was based on the application of the dynamic time warping (DTW) algorithm with Euclidean distance in order to compare each test sample to each pattern sample [Davis, 80] [Paliwal, 82a, 82b and 84]. ESPS software [ESPS, 93] was used to compute, store and plot the parameters. Table II summarizes the results of these experiments.

Table II. Recognition rates obtained in the 5-fold cross-validation experiments.

Parameter	Vector size	Preemphasis	BLT coefficient	% Recognition
BF	21	NO	-----	99.2
BFB	21	YES	-----	99.2
BFBCEP	12	NO	-----	99.9
BFBCEP	12	YES	-----	99.9
FFTCEP	12	NO	-----	96.6
FFTCEP	12	NO	0.4, 0.5, 0.6, 0.7, 0.8	98.1, 98.4, 98.3, 97.6, 97.9
FFTCEP	12	YES	-----	96.2
FFTCEP	12	YES	0.4, 0.5, 0.6, 0.7, 0.8	98.1, 98.3, 98.4, 97.5, 97.5
LPCEP	12	NO	-----	95.8
LPCEP	12	NO	0.4, 0.5, 0.6, 0.7, 0.8	97.7, 98.1, 98.9, 99.1, 98.9
LPCEP	12	YES	-----	95.9
LPCEP	12	YES	0.4, 0.5, 0.6, 0.7, 0.8	98.6, 98.9, 99.3, 99.6, 99.2
RC	14	NO	-----	93.2
RC	14	YES	-----	94.4
LAR	14	NO	-----	96.3
LAR	14	YES	-----	94.5

Figure 1a shows the recognition rates obtained through the parametric representations LPCEP and FFTCEP described in the previous section and the different values of the BLT coefficient. From this figure, we can conclude that the optimal value of the this coefficient depends on the analysis method. A value of 0.56 matches the Bark scale [Rodriguez, 94] but different optimal values were found for both sets of parameters. Figure 1b shows the recognition rates for the same experiments when the best BLT coefficient was chosen for different vector sizes. The BFB coefficient vector was no longer considered because of the large computation required in the recognition task. The RC and LAR parameters provided poor recognition rates and thus are not represented. In Figure 1b, several parameters appear with very high recognition rates. Most of them are also very similar and it is not easy to evaluate their performances.

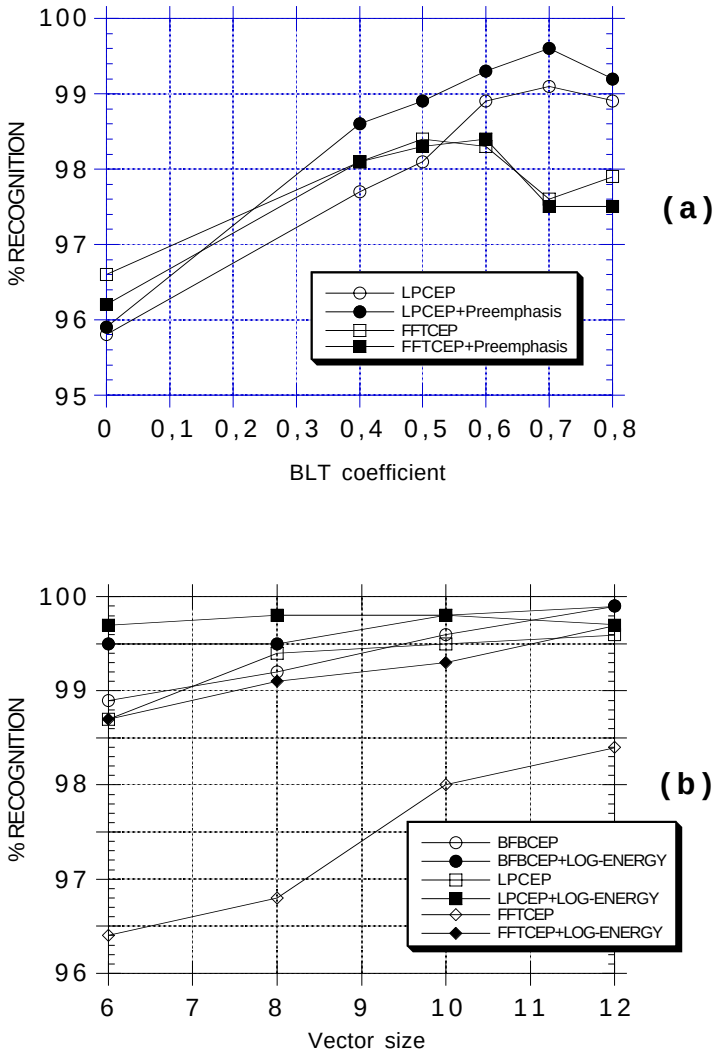

Figure 1. Recognition rates obtained through the first series of experiments: (a) LPCEP and FFTCEP with different values of the BLT coefficient, and (b) BFBCEP, LPCEP and FFTCEP with different vector sizes and optimal choices for both the BLT coefficient and preemphasis.

In order to increase the difficulty of the task, a second series of experiments was carried out. In each partition the training size was 200 utterances and the test set was 800 utterances, yielding an effective test set of 4000 utterances. Only BFBCEP, LPCEP and FFTCEP were considered in this case. Figure 2 shows the recognition rates for these experiments for different values of the BLT coefficient (Figure 2a) and different vector sizes (Figure 2b). BFBCEP provided the best system performance when the vector size was 12. However, LPCEP with preemphasis and bilinear transformation provided higher recognition rates when the vector size was 6 and 8. Adding the log-energy to the vector of parameters improved the recognition rates significantly, especially for FFTCEP and LPCCEP. LPCCEP+log-energy provided the best recognition rates with vector sizes 6, 8 and 10, and only slightly lower rates than BFBCEP+log-energy with vector size 12.

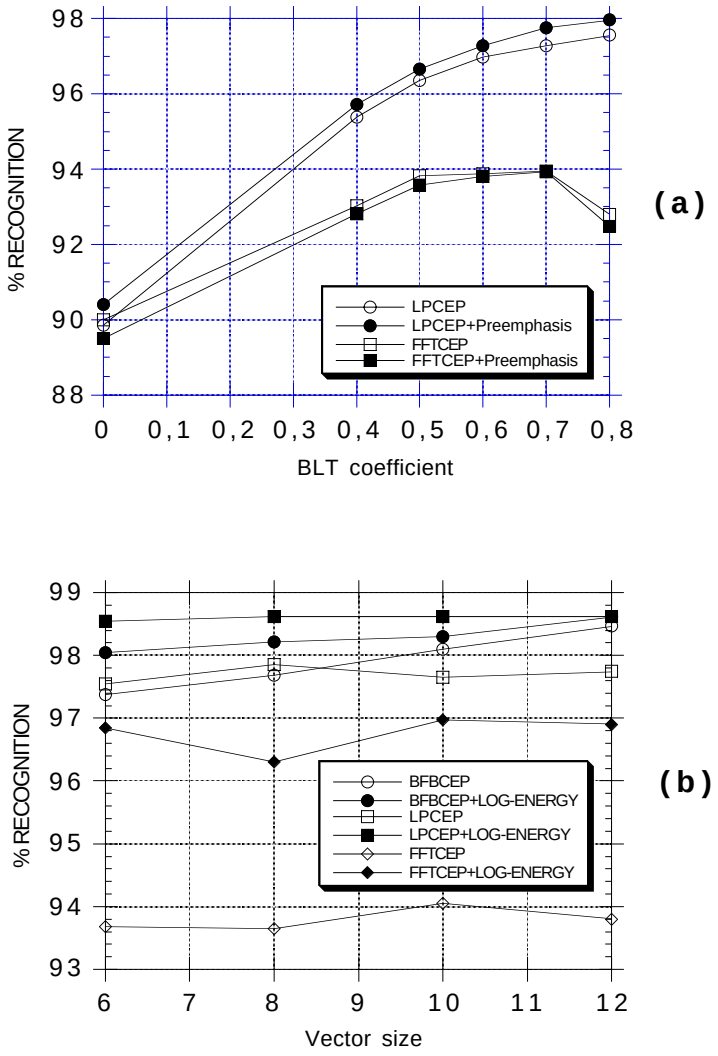

Figure 2. Recognition rates obtained through the second series of experiments: (a) LPCEP and FFTCEP with different values of the BLT coefficient, and (b) BFBCEP, LPCEP and FFTCEP with different vector sizes and optimal choices for both the BLT coefficient and preemphasis.

The statistical behaviour of the parameters should also be considered when choosing among several spectral analysis procedures, because of its influence in the modelling approach [Nocerino, 85][Shikano, 86][Tohkura, 87]. The computation of the variance for each component was also needed to re-scale the energy before adding it to the vector of parameters. Figure 3 represents the mean and the variance of each component of BFBCEP, LPCEP and FFTCEP. In all the cases the variance decreased quickly to a low value, and the mean tended to be slightly negative. Only a few components, approximately six, had a significant variance. This could explain the lack of improvement introduced in the recognition rates by the last components of the vectors. It seems that the higher the variances of the last components, the larger the improvement introduced in the recognition rates. On the other hand, first components are primarily affected by the analysis method. This could explain the different recognition rates obtained by each parameter vector.

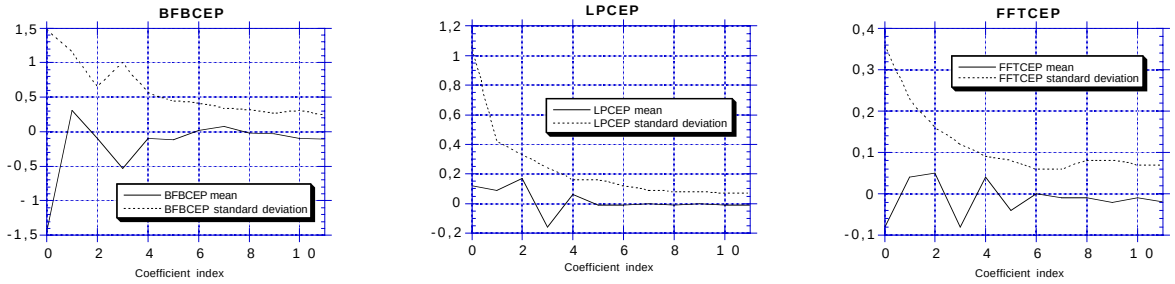

Figure 3. Mean and standard deviation of the vectors BFBCEP, LPCEP and FFTCEP.

Concluding remarks-.

The experiments carried out in this work confirm the cepstral coefficients as the best choice to represent speech signals in speech recognition tasks. These coefficients were obtained through FT and LP analysis, with the first method leading to better system performance. Introducing the Bark scale in the computation of the filter bank coefficients and the bilinear transformation in the computation of both the LP and the FFT based cepstral coefficients increased the recognition rates, but particularly in the first case, where BFBCEP provided the best system performance.

Preemphasis improved the performance of LPCEP but produced lower rates in Bark-scaled filter bank parameters, and did not affect FFTCEP. The choice of the best parametric representation depends on the vector length (L): LPCEP would be clearly the choice for L=6 and L=8, and BFBCEP the choice for L=10 and L=12. Adding log-energy, with variance equal to the maximum variance of the vector of parameters, also significantly improved the recognition rates.

Some experiments with the first and second derivatives should be done to complete this work. The components could be weighted with *liftering* windows [Tohkura, 87] [Segura, 91], and the performance tested for several lengths of the vector. At present we are testing some of these parameters in acoustic-phonetic decoding experiments, where sublexical units are modelled using hidden Markov models.

Acknowledgements-.

The corpus used in this work was supplied by the Group of Pattern Recognition and Artificial Intelligence of the Universidad Politécnica de Valencia (Spain). The authors wish to thank all of them, especially J. M. Benedí and E. Vidal for their help and suggestions.

References-.

[Bellegarda, 90] J. R. Bellegarda and D. Nahamoo. “Tied Mixture Continuous Parameter Modelling for Speech Recognition”. *IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 38, No. 12, pp. 2033-2045, 1990*

[Davis, 80] S.B. Davis, P. Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *IEEE Trans. on ASSP, vol. 28, n. 4, pp. 357-366. August 1980.*

[ESPS, 93] "Entropic Signal Processing System Reference Manual". Version 5.0. *Entropic Research Laboratory. Washington DC. August 1993.*

[Huang, 93] X. D. Huang, H. W. Hon, M. Y. Hwang and K. F. Lee. “A comparative study of discrete, semicontinuous and continuous hidden Markov models”. *Computer Speech and Language Vol. 7, pp. 359-368, 1993.*

[Lee, 89] K.F. Lee. "Automatic Speech Recognition. The development of the SPHINX System". *Kluwer Academic Publishers. 1989.*

[Levinson, 85] S.E. Levinson. "A Unified theory of composite Pattern Analysis for Automatic Speech recognition". *Computer Speech Processing, pp. 243-275. Ed. F. Fallside, W.A. Woods. Prentice Hall.*

[Nocerino, 85] N. Nocerino, F.K. Soong, L.R. Rabiner, D.H. Klatt. "Comparative study of several distortion measures for speech recognition". *Speech Communication, vol. 4, n. 4, pp. 317-331. December 1985.*

[Oppenheim, 72] A.V. Oppenheim, D.H. Johnson. "Discrete representation of signals". *Proc. of the IEEE, vol. 60, n. 6, pp. 681-691. June 1972.*

[Paliwal, 82a] K.K. Paliwal, P.V.S. Rao. "Evaluation of various linear prediction parametric representations in vowel recognition". *Signal Processing, vol. 4, n. 4, pp. 323-327. July 1982.*

[Paliwal, 82b] K.K. Paliwal. "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition". *Speech Communication, vol. 1, n. 2, pp. 151-154. August 1982.*

[Paliwal, 84] K.K. Paliwal. "Effect of preemphasis in vowel recognition performance". *Speech Communication. vol. 3, n. 1, pp. 101-106. April 1984.*

[Raudys, 91] S. J. Raudys and A. K. Jain, “Small Sample Effects in Statistical Pattern Recognition: Recommendations for Practitioners and Open Problems,” *IEEE Trans. on PAMI, vol. 13, n3, pp. 252-263, 1991.*

[Rodríguez, 94] L.J. Rodríguez. "Estudio comparativo de varias representaciones paramétricas para reconocimiento automático del habla". *Informe de Investigación DEE-I/2/94. Grupo de Reconocimiento Automático del Habla. Departamento de Electricidad y Electrónica. Universidad del País Vasco. Noviembre 1994.*

[Segura, 91] J.C. Segura. "Modelos de Markov con cuantificación dependiente para reconocimiento de voz". *Tesis Doctoral. Departamento de Electrónica y Tecnología de Computadores. Universidad de Granada. Noviembre 1991.*

[Shikano, 86] K. Shikano. "Evaluation of LPC spectral matching measures for phonetic unit recognition". *Technical Report CMU-CS-86-108. Computer Science Department. Carnegie-Mellon University. February 1986.*

[Tohkura, 87] Y. Tohkura. "A weighted cepstral distance measure for speech recognition". *IEEE Trans. on ASSP, vol. 35, n. 10, pp. 1414-1422. October 1987.*

[Zwicker, 81] E. Zwicker, R. Feldtkeller. "Psychoacoustique. L'oreille, recépteur d'information". *Masson Ed. París, 1981.*