Feature Extraction from Time-Frequency matrices for Robust Speech Recognition

José C. Segura[1], *M. Carmen Benítez*[1,2], *Ángel de la Torre*[1], *Antonio J. Rubio*[1]

[1]Dept. of Electronics and Computer Architecture, University of Granada, Spain
[2] International Computer Science Institute Berkeley, CA, USA
segura@ugr.es

Abstract

In this paper we present a study about time-frequency distribution of acoustic-phonetic information for the Spanish language. This is based on a large Spanish database automatically labeled, and we conclude that results are similar to those obtained for hand-labeled english databases. We use bidimensional LDA [1] to extract discriminant features in time-frequency domain (TF) that are more robust in noise than the standard ones based on MFCC and time derivatives. We show that TF domain and its corresponding transformed domain (CTM) are equivalent from the point of view of LDA analysis and use this fact to reduce the dimensionality of the problem. Finally, cascade unidimensional LDA (CLDA) is applied first in frequency and then in time. This gives better estimates of projection vectors and better recognition performance. The proposed techniques are evaluated in a connected digit recognition task. Utterances have been artificially corrupted with additive real noises.

1. Introduction

Most speech recognition systems use log filter bank energies (FBE) from short frames (30 ms) for the acoustic representation of speech. The temporal sequence of log FBE $\mathbf{S}_t$ is usually modeled using continuous observations HMM with diagonal covariance matrices. As FBE's are highly correlated, it is necessary to use a decorrelation transform (usually DCT or KLT). HMM modeling also assumes independent temporal observations, and this makes difficult to represent the parmeter dynamics. To overcome this, it is common to augment the parameter vector with its time derivatives (delta parameters).

To improve the robustness of the speech parameterization, frequency an time correlations have been used to define new parameterization techniques. In this way, alternatives for the frequency transform (DCT) and time representation (static plus delta parameters) have been proposed that result in a more robust set of parameters. Several authors have proposed alternative transforms [2], [3], [4], [5] that offer better performance in noisy conditions.

Yang et al [6] have used the concept of mutual information to show the time-frequency distribution of mutual information between log FBE and phonetic classes (acoustic-phonetic information) for the english language. The global time-frequency correlation suggest the use of a joint technique to extract the acoustic features.

The main objective of this work is the design of linear time-frequency transforms that extract the maximum amount of phonetic information from acoustic observations. This will be done by means of a bidimensional LDA analysis of TF domain. As the resulting features will focus on most relevant parts of this domain, a more robust behavior is expected.

The rest of this paper is organized as follows. In section 2 we show results on time-frequency distribution of acoustic-phonetic information for the Spanish language. In section 3 we present the bidimensional LDA analysis of TF domain and give some comparative results on a noisy continuous digit recognition task. Section 4 is devoted to LDA analysis on CTM domain and we also present there some comparative results between CTM-LDA and column selection of CTM matrices. Cascaded LDA (CLDA) is evaluated in section 5 and in section 6 we summarize the results and conclusions of this work.

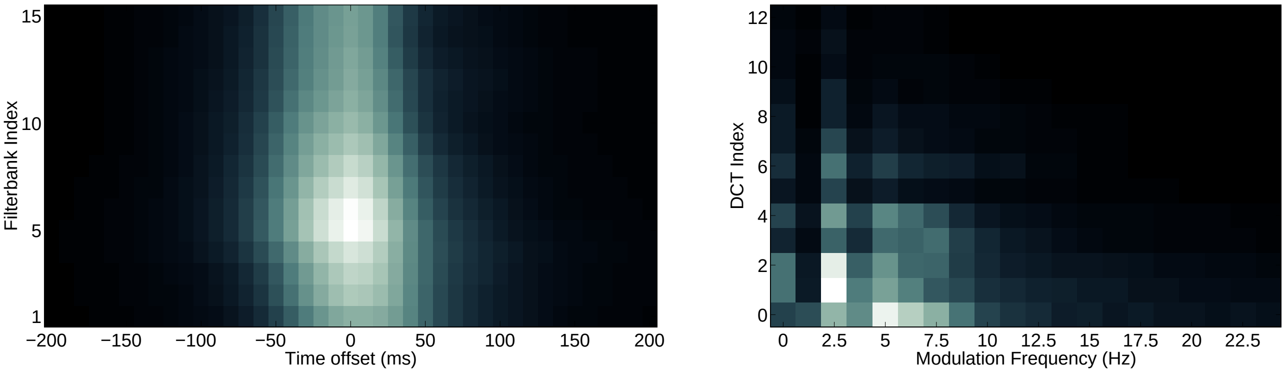
2. Distribution of acoustic-phonetic information for the Spanish language

The distribution of acoustic-phonetic information for the English language is investigated in [6]. In this section we present similar results for the Spanish language.

Time-frequency information distribution is represented by the mutual information between log FBE and phonetic classes. We have used a 8 KHz down sampled version of ALBAYZIN [7], [8] database. The training partition is a phonetically balanced[1] recording of about 3 hours of speech. It contains 4.000 utterances from 160 speakers. This database was automatically labeled using forced alignment with a set of 24 phonemes using HTK and HMM context independent phone models with three emitting states. From this alignments, each speech frame is labeled and associated with a time-frequency matrix $\mathbf{S} = \{S(n,\tau)\}$ containing log FBE from $N = 15$ mel frequency spaced filters covering the entire frequency range (0-4 KHz) and delays $\tau = -\frac{T}{2} \cdots \frac{T}{2}$ with $\tau = 0$ for the current labeled frame. Considering a time-span of $\pm 200ms$ around the current frame, at a frame-rate of 100 Hz the resulting matrices are of size $15 \times 41 = 615$. Using histograms like in [6], mutual information of time-frequency matrix components and phoneme classes have been obtained and are depicted in figure 1(a) as gray level images ranging from 0 (black) to 0.8 bits (white). From this representation it can be observed that mutual information reaches its maximum value for filter bank index $n = 5$ and time delay $\tau = 0$, and is spread along both time and frequency axes. Spread in time is almost symmetric in an interval of about $250ms$ around the center frame (information out of this interval is smaller than 0.04 bits). This domain will be referred in the following as the TF domain.

Several others domains have been alternatively proposed as

[1] Statistical distribution of allophones is close to that of Spanish language



(a) TF domain

(b) CTM domain

Figure 1: *Acoustic-phonetic information distribution.*

the two dimensional modulation spectrum (2D-MS) [2] or Cepstral Time Matrix (CTM) [9]. In both cases frequency axis is linearly transformed into cepstral domain with a DCT and time axis is transformed into modulation frequency domain by a DFT or a DCT respectively. As both domains are similar we only present here results about CTM. CTM is obtained from TF with a two dimensional DCT transform. Information distribution in this domain is obtained in a way similar of that used for TF and is showed in figure 1(b). Only modulation frequencies up to 25 Hz and cepstral coefficients up to 12 are used in this analysis.

In CTM domain, the mutual information appears concentrated in a few number of lower index rows (low quefrency) and a few number of lower index columns (low modulation frequency) of CTM matrix. In fact, columns with associated modulation frequency out of 2-16 Hz range have very little information content. This suggest that time filtering cepstral coefficients to enhance this modulation frequency range could help to reduce the mismatch in noisy conditions. It is also evident from the information distribution that the adequate frequency range is different for each cepstral coefficient. Lower order ones have a broader useful modulation frequency range than higher order ones. This is due to the fact that lower order cepstral coefficients are smoother by the low pass filtering effect of the frequency axis transform.

This interaction of transforms in time and frequency has also been pointed by Nadeu et al [2] and supported with experimental results. As a consequence, optimal transforms must jointly consider time and frequency. Although it is possible to search for general transforms, in this work we restrict ourselves to linear ones based on LDA.

3. Feature extraction

The main objective of the desired feature extraction technique must be to emphasize the region of maximum mutual information, reducing the influence of low information parts. This way the mismatch in noisy conditions will be reduced. A direct approach based on CTM domain has been proposed by Milner and Vaseghi [9], [10] and correspond to retain a subset of adjacent columns of CTM excluding the first one that corresponds to zero modulation frequency. Nadeu et al [2] have proposed the use of frequency and time filters (tiffing) to emphasize the most robust part of the 2D-MS. Hermansky [3], Avendano [4], van Vuuren [5] and others have proposed to use LDA for both time and frequency. In this section we explore the performance of a direct bidimensional LDA analysis to perform the feature extraction.

3.1. LDA in TF domain

Denoting by $\mathbf{S} = \{S(n,\tau)\}$ a matrix in the TF domain, a linear transform $\mathbf{H} = \{H(k,n,\tau)\}$ which maps it into a vector $\mathbf{F} = \{F(k)\}$ in the feature space (FS) of dimension K whose components can be written as

$$F(k) = \sum_{n=1}^{N} \sum_{\tau=-T/2}^{T/2} H(k,n,\tau) S(n,\tau) \qquad \forall k = 1..K \qquad (1)$$

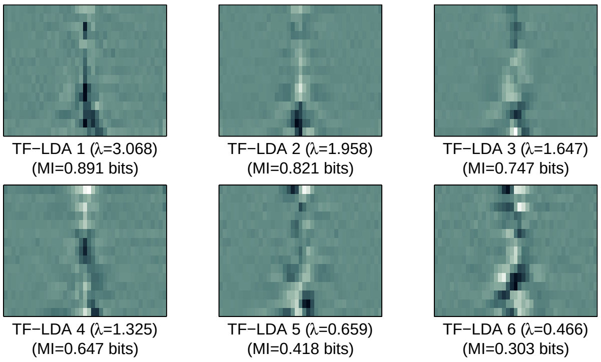
Without loss of generality, we can convert the two dimensional TF space in a one dimensional one by means of and adequate index remapping

$$\begin{aligned} F(k) &= \sum_{v=1}^{(T+1)N} \underline{H}(k,v)\underline{S}(v) \qquad \forall k = 1..K && (2)\\ \underline{S}(v) &= S(n,\tau) && (3)\\ \underline{H}(k,v) &= H(k,n,\tau) \ / \ v = (\tau + T/2)N + n && (4) \end{aligned}$$

or in matrix notation $\mathbf{F} = \underline{\mathbf{H}}\,\underline{\mathbf{S}}$. With this definitions, bidimensional analysis of $\mathbf{S}$ can be done by means of unidimensional LDA of an augmented vector $\underline{\mathbf{S}}$. The optimization criterion used is to maximize $\mathbf{tr}\left\{\mathbf{V_w^{-1}V_b}\right\}$, where $\mathbf{V_w}$ and $\mathbf{V_b}$ are the within and between class covariance matrices [1] of the elements of $\underline{\mathbf{S}}$. The transform matrix $\underline{\mathbf{H}}$ have as rows the solutions of the generalized eigenvalue problem $\mathbf{V_w \Phi} = \mathbf{V_b \Phi \Lambda}$

Following this approach, an LDA analysis of TF domain has been performed on ALBAYZIN phonetic training partition. Spectral vectors have been derived as indicated in section 2. We have considered 72 classes corresponding to the three states[2] of each phone model (excluding the silence model). As a result we obtain a set of 71 eigenvectors. The first 6 of them with greater associated eigenvalue are shown in figure 2 once rearranged into matrix form. We also show the associated eigenvalue and mutual information obtained as described in section 2 for log FBE. Effectiveness of LDA analysis is demonstrated considering the information extracted by each eigenvector. The best TF-LDA feature extracts 0.89 bits of information while only 0.70 bits are extracted by the best MFCC_D_A.

[2] Preliminary experiments have shown that state segmentation performs better than phone segmentation.



TF−LDA 1 (λ=3.068) (MI=0.891 bits)

TF−LDA 2 (λ=1.958) (MI=0.821 bits)

TF−LDA 3 (λ=1.647) (MI=0.747 bits)

TF−LDA 4 (λ=1.325) (MI=0.647 bits)

TF−LDA 5 (λ=0.659) (MI=0.418 bits)

TF−LDA 6 (λ=0.466) (MI=0.303 bits)

Figure 2: *First 6 eigenvectors in TF domain.*

3.2. Performance in additive noise

To test the proposed technique, a set of recognition experiments have been carried out in a continuous digit recognition task. The database used for this purpose consist of 600 strings containing 4800 Spanish digits uttered by 20 adult speakers (10 male and 10 female). This database is split into two partitions of 300 strings for training and 300 for test. Recordings have been made at a sampling frequency of 8 KHz in PCM format with 16 bits. The mean SNR is 43 dB. Training is always performed in clean condition. For recognition, an artificially corrupted version of test partition is built by mixing the four noises from AURORA v2.0 database test set a (SUBWAY, BABBLE, CAR and EXHIBITION) at SNR's between 20 dB and -5 dB.

Recognition is performed using HTK with a 16 emitting states 5 mixtures continuous HMM model for each digit (silence model has only 3 emitting states). Basic parameterization consist of log FBE from 15 mel spaced filters as implemented in HTK. The length of the analysis window is 30 ms an the frame rate is 100 Hz. In the reference test, MFCC_0_D_A is used with 13 cepstral coefficients (including C_0) augmented with delta and acceleration coefficients with regression lengths of 3 and 2 respectively.

Table 1 shows the word accuracy averaged over the 4 noise conditions. Columns two and three correspond to standard MFCC_0_D_A and TF-LDA with same number of features (39). Effectiveness of bidimensional LDA is stated as the only situation in which it performs worse than the standard is the clean condition (only 0.04% of absolute degradation). In any other situation, TF-LDA performs better, specially at 15 dB where the relative error reduction is of 51.88%.

4. Feature selection in CTM domain

As described in the previous section, an alternative approach is to perform the feature extraction in the transformed CTM domain. The first approach we have evaluated is to use a submatrix selection criterion. We have used the same basic parameterization as in the former case, and then we have generated CTM matrices with 9×4 and 13×3 components obtained from TF matrices in a temporal interval of 15 frames. In the frequency axis, a DCT is applied and we keep the cepstral coefficients $C_F(0)-C_F(8)$ or $C_F(0)-C_F(12)$, and in time we apply DCT and the coefficients $C_T(1)-C_T(4)$ or $C_T(1)-C_T(3)$ are kept, respectively. This amounts to 36 or 39 features, respectively.

The results obtained over the same evaluation database are shown in table 1. A comparison with the previously presented results shows that, with the exception of the clean conditions, the TF-LDA method outperforms CTM. The improvement is due to the fact that CTM is implemented using a temporal interval of 150 ms, while TF-LDA extracts the features from a temporal interval of 410 ms, which allow the inclusion of more discriminative information.

SNR	MFCC_D_A	TF-LDA	CTM-9x4	CTM-13x3
clean	99.42%	99.38%	99.50%	99.17%
20dB	90.35%	96.18%	95.26%	95.36%
15dB	73.15%	87.04%	85.64%	85.61%
10dB	51.53%	66.38%	60.72%	62.80%
5dB	27.54%	40.36%	29.37%	32.96%
0dB	13.48%	22.50%	11.90%	15.06%
-5dB	9.26%	13.65%	7.39%	8.29%

Table 1: MFCC_D_A, TF-LDA and CTM performance

4.1. LDA in CTM domain

If the temporal range considered for CTM is extended to 410 ms, CTM and TF are domains related by the bidimensional DCT transform. In this situation, obtaining a simple feature selection criterion is difficult, due to the high number of involved components. Preliminary tests for feature selection in order to reduce to 39 the number of components based on a criterion of maximum mutual information did not provide proper results. However, the application of LDA in this domain generates a set of eigenvalues and eigenvectors that are, in theory, identical to those obtained by TF-LDA, since CTM and TF are related by an orthonormal transform.

In order to evaluate CTM-LDA, we have considered the transformation of the matrices $\mathbf{S}$ (in the TF domain) to $\mathbf{C}$ in the CTM domain through a bidimensional DCT. Only the rows of $\mathbf{C}$ corresponding to the first 13 cepstrum (including C_0) and the columns corresponding to the modulation frequencies between 0 and 25 Hz (20 first columns) are preserved. After an index reassignment, the resulting vector dimensionality is $13 \times 20 = 260$ (significantly lower than in the case of TF-LDA, where the dimensionality is $15 \times 41 = 615$).

The eigenvalues obtained for TF-LDA and CTM-LDA are roughly similar, with the exception of the first one, for which TF-LDA provides a value (3.068) slightly higher than that for CTM-LDA (2.988). This difference is due to the elimination of the CTM columns for modulation frequencies over 25 Hz.

With respect to the performance in noise conditions, table 2 shows the results obtained for CTM-LDA (sing 39 features). The results are similar as those obtained for TF-LDA. We think the slight differences in favor of CTM-LDA are due to the better estimation of the eigenvectors thanks to the dimensionality reduction when the CTM columns are truncated (which introduces a 25 Hz low-pass filtering).

5. Cascade of unidimensional LDA

In spite of the use of more than a million of frames for the estimation of the covariance matrices for the LDA analysis, the obtained eigenvalues are rather noisy, as observed in figure 2. The imprecise estimation of eigenvectors introduce imprecisions in the features obtained as a projection over them and this degrades the recognition performance. This situation has been previously reported by Kajarekar et al [11] when bidimiensional LDA anal-

SNR	CTM-LDA	TF-CLDA
clean	99.33%	98.75%
20dB	96.16%	97.09%
15dB	87.70%	91.14%
10dB	68.18%	72.18%
5dB	40.79%	41.72%
0dB	21.51%	17.86%
-5dB	12.39%	10.56%

Table 2: CTM-LDA and TF-CLDA performance

ysis is applied to the OGI Stories database, using vectors with 1515 components.

An alternative solution to avoid this problem is the independent design of the LDA transformations for time and frequency. However, this approach does not allow to use the time-frequency cross correlations. In order to preserve these correlations, we apply a cascade LDA (CLDA). Firstly a LDA transformation is estimated for log-FBE components in the current frame (delay $\tau = 0$), and the 13 first eigenvectors are considered to project the log-FBE in the new component set. Secondly, a temporal LDA analysis is performed, independently for each of the new components, and the first 3 eigenvectors are considered. The combination of the 13 frequency-domain eigenvectors and the correspondent 3 time-domain eigenvectors provides a feature set with 39 parameters.

Figure 3 shows the first 6 eigenvectors obtained with cascade LDA approach in TF domain (TF-CLDA). The comparison of these eigenvectors to the ones obtained by bidimensional LDA over TF (TF-LDA figure 2) shows the new ones are less noisy. The TF-CLDA approach provides better recognition performance in the range from 20 dB to 5 dB as observed in the last column of table 2.

6. Conclusion

This work presents a study about the distribution of mutual information between spectral observations and phonetic classes for Spanish, obtained by means of an automatic segmentation of a large Spanish phonetic database. The obtained distribution is qualitatively similar to that previously reported for English using hand-labeled databases.

Making use of a bidimensional LDA analysis as a feature extraction method we have derived a set of robust features. The recognition experiments under additive noise conditions show that this parameterization is more robust than the standard one based on static plus dynamic cepstral coefficients.

We have also studied the transformed domain CTM, showing that it is equivalent to the TF domain as a start point for LDA. The information distribution in the CTM domain is used in order to perform a reduction of the space dimensionality previous to the LDA analysis. This dimensionality reduction leads to a better estimation of the discriminative transformation, providing better recognition results.

Alternatively to the feature reduction in the whole cepstral-time vectorial space, LDA has been also applied in cascade to each dimension of TF domain. Even though the joint LDA is theoretically a better approach, the cascade LDA method provides better recognition results in practical applications, mainly because of the limitations of the available data for the analysis. However, when the mismatch between training and test conditions is more important (very low SNR) joint LDA is better than cascade LDA.

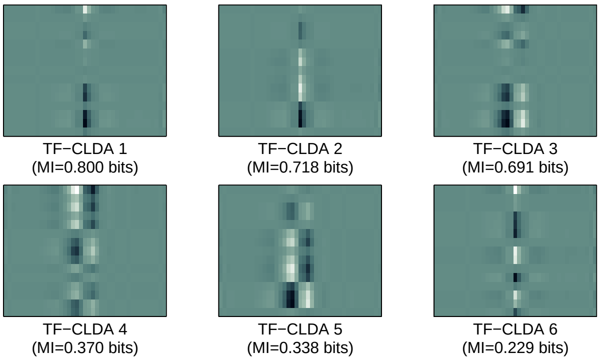


Figure 3: *Eigenvectors TF-CLDA.*

7. Acknowledgments

This work has been supported by the Spanish Government under CICYT project TIC99-0583 and ICSI grant “Robust features extraction for speech recognition in GSM mobile telephony environment”

8. References

[1] K. Fukunaga, “Statistical pattern Recognition”, Academic Press, San Diego, 1990.

[2] C. Nadeu, D. Macho, J. Hernando, “Time and frequency filtering of filter-bank energies for robust HMM speech recognition”, Speech Communication, Vol. 34(1), pp 93-114, 2001

[3] H. Hermansky and N. Malayath, “Spectral basis functions for discriminat analysis”, Proc. ICSL, Sydney, 1998

[4] C. Avendano, S. van Vuuren, H. Hermansky, “Data-Based RASTA-like filter design for channel normalization in ASR”, Proc. ICSLP’96, Philadelphia, 1996, pp 2087-2090

[5] S. van Vuuren and H. Hermansky, “Data-Driven Design of Rasta-Like Filters”, Proc. EUROSPEECH’97, 1997

[6] H.H. Yang, S. van Vuuren, S. Sharma and H. Hermansky, “Relevance of Time-Frequency Features for Phonetic and Speaker-Channel Classification”, Vol. 31(1) pp 35-50, 2001

[7] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J.M. Pardo and A. Rubio, “Development of Spanish Corpora for Speech Research (ALBAYZIN)”, Proc. of the Workshop on International Cooperation and Standardization of Speech Databases and Speech I/O Assesment Methods. Chiavari 26-28 September, 1991 (Italy).

[8] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, L.B. Mario, C. Nadeu, “ALBAYZIN speech data base: Design of the phonetic cospus”, Proc. EUROSPEECH’93, pp. 175-8, Berlin, 1993

[9] B.P. Milner and S.V. Vaseghi, “An analysis of cepstral-time matrices for noise and channel robust speech recognition”, Proc. EUROSPEECH’95, pp 519-552, 1995

[10] B.P. Milner, “Cepstral-time matrices and LDA for improved connected digit and sub-word recognition accuracy”, Proc. EuROSPEECH’97, 1997

[11] Sachin S. Kajarekar, B. Yegnanarayana and H. Hermansky, “A Study of Two Dimensional Linear Discrimination for ASR”, submitted to ICASSP’2001.