Normalization of the inter-frame information using smoothing filtering

L. García , J.C. Segura , C. Benítez, J. Ramírez, A. de la Torre

Departamento de Teoría de la Señal Telemática y Comunicaciones
University of Granada, Spain
{luzgm,segura,carmen,javierrp,atv}@ugr.es

Abstract

A filter that introduces inter-frame information into the voice features set is proposed in this paper. The filter adds the autocorrelations of the cepstral coefficients to the set of characteristics used for training and recognition. Those autocorrelations should not depend on the environment conditions. Because they should only depend on the information to recognize, a normalization of that inter-frame information is convenient. The filter defined implements this normalization by transforming the autocorrelations into a normalized domain defined with clean adaptation data. This temporal processing of the features is added to the Histogram Equalization of the cepstral coefficients *(HEQ)* used to normalize the MFCCs. An analysis is done about the most effective domain (original MFCCS or equalized MFCCs) on which the temporal processing should be executed. Performance results for the proposed algorithm are presented for AURORA2 and AURORA4 databases.

Index Terms: robust speech recognition, temporal filtering, frames correlation

1. Introduction

The analysis of the inter-frame information contained in the speech contributes valuably to the recognition process by adding univocal characteristics to the data under process. Still only a limited number of algorithms for features extraction take into account this temporal information. In the case of the MFCC parameters, the basic temporal processing done is the use of the time derivative parameters delta and delta-delta cepstra. Other techniques to capture the inter-frame information are the ones derived from the RASTA filtering [1, 2, 3] implemented alone, or optimized using LDA or PCA analysis [4].

Like the rest of characteristics that compose the voice features, this inter-frame information is sensitive to environment mismatches such as background noise and channel distortion. These mismatches degrade the recognition performance in adverse conditions. For this reason it is desirable to also normalize the temporal information like it is done for the rest of parameters. It also should be invariant to environmental changes. Histogram Equalization HEQ has proven to be an optimum normalization technique [5, 6, 7] that outperforms the Cepstral Mean and Variance Normalization by removing non-linear irrelevant information of the cepstral coefficients. The application of HEQ also to the delta and delta-delta cepstra has been proposed and studied in [8, 9]. The dependent or independent equalization of those coefficients has been analyzed obtaining an optimal normalization technique: a feedback calculation is performed using the dynamic cepstra to equalize the static cepstra, and then recalculating again the dynamic ones with the latest equalized static coefficients.

The motivation of our work is to propose a different technique to include and normalize temporal information. We propose a temporal smoothing filter for the voice features added to the HEQ features normalization. This filter searches two objectives. The first one is to introduce inter-frame information into the features set by taking into account the inter-frame correlation of each cepstral coefficient. The second objective is to normalize that inter-frame information by defining a linear transformation applied to the inter-frame information of the test and training data that are moved into a common domain defined with clean adaptation data. In order to accomplish this, the rest of the paper is organized as follows. In section 2 the temporal smoothing filter proposed is described. In section 3 the experiments and results are presented. Conclusions and future work are given in section 4.

2. Temporal smoothing filter

2.1. Filter definition

A simple ARMA filter is proposed to restore the temporal autocorrelation structure of each cepstral coefficient. The filter can be obtained as the cascade of two filters. The first one is a whitening filter that removes the temporal correlation structure of the input data; while the second filter restores the desired temporal correlations.

Given a temporal sequence of observations of a particular cepstral coefficient $x(n)$, a whitening filter is designed to transform it to an uncorrelated sequence $u(n)$. This can be done under a linear prediction approach. Let us denote $A(z)$ the predictor,

$$U(z) = A(z)X(z) \tag{1}$$

Once the whitened sequence $u(n)$ is obtained, a second filter is used to restore the desired autocorrelation structure. This second filter is also derived under a LPC approach.

$$Y(z) = \frac{U(z)}{B(z)} \tag{2}$$

Finally, the filter is obtained as the cascade of the two previous filters

$$Y(z) = \frac{A(z)}{B(z)}X(z) = H(z)X(z) \tag{3}$$

which results in an ARMA filter.

The coefficients of the whitening filter (the first one) are derived from the autocorrelation coefficients of the actual utterance. The coefficients of the second filter are obtained from an estimation of the autocorrelations of reference data (i.e. clean training or adaptation data).

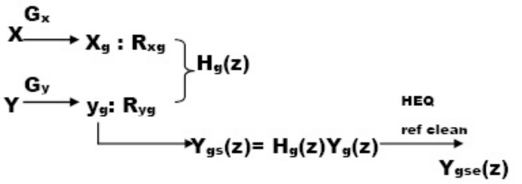

Figure 1: Temporal smoothing done in the Gaussian domain

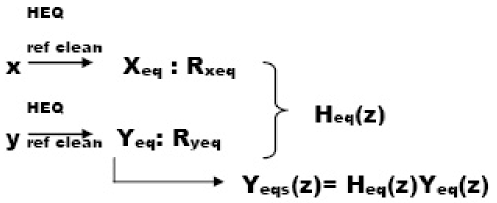

Figure 2: Temporal smoothing done in the equalized domain

2.2. Filter location

Using *ys*, the smoothed version of *y*, implies applying a linear transformation to the features that consists of a set of scalings which make the temporal autocorrelations maintain the same proportions and ratios than in the reference domain defined with clean adaptation data. The effect of this scaling will be the elimination of inter-frame information distortions originated by a test environment different from the training one. The domain in which this temporal smoothing is done has an effect on the improvement achieved. It is desirable to locate the filter in the domain where the essential temporal information is more clearly separated from the irrelevant one that should be removed. As we mentioned in the introduction about normalization, applying HEQ to the front-end features means transforming them non-linearly to suit a reference Cumulative Distribution Function (CDF). With this operation, the training and test MFCC parameters are moved to a domain which is more robust against the environment peculiarities[5]. Two reference CDF used for the equalization have proved to work optimally: the clean data CDF, and a Gaussian CDF. According to this, three possible scenarios to implement the filter transformation have been analyzed. They are depicted in figures 1, 2 and 3

In figure 1 the autocorrelations of the clean data R_{xg} and the test sentence (R_{yg}) are calculated once the features have been equalized to a reference Gaussian CDF. The filter is defined in a Gaussian domain and after normalizing the inter-frame info via the filter, the *smoothed* features are equalized using a clean reference CDF. In figure 2, the features are equalized directly using a clean reference CDF, and then the correlations for the clean and test data are calculated (R_{xeq} and R_{yeq}). The smoothing filter is defined and applied in the equalized domain. For the last scenario shown in 3, the temporal filtering is done calculating the autocorrelations (R_x) and (R_y) in the original domain. The smoothed parameters are later equalized using a clean reference CDF.

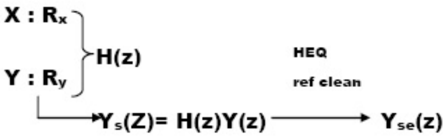

Figure 3: Temporal smoothing done in the original domain

3. Experiments and results

3.1. Description of the test environment

The recognition system used is based on continuous cross-word triphone models with 3 tied states and a mixture of 6 Gaussians per state. Training and recognition are done using the HMM Tool Kit (HTK) software. The language model is the standard bigram for the WSJ0 and AURORA2 tasks. A feature vector of 13 cepstral coefficients is used as the basic parameterization, using Co instead of the logarithmic energy and spectral mean substraction. This basic features vector is augmented with first and second order regressions yielding a final 39 components features vector.

The algorithm proposed in this paper has been tested for AURORA4 and AURORA2 databases following the standard clean training tests. All the procedures for training and recognition are identical to the reference experiments, with the exception of the front-end that includes the histogram equalization and the temporal smoothing under study. The parameters of the reference distribution used for the equalization have been obtained by averaging over the whole clean training set of utterances. Training and test utterances have been temporarily smoothed and normalized using a set of clean adaptation data to calculate the reference autocorrelation of the smoothing filter. The whole process of non linear and linear transformations has been done before computing the regressions. This decision is supported by experimental tests that pointed out a 3.5 of increase in the WER if the regressions were computed before the equalization and temporal smoothing.

For comparison purposes, three more experiments have been conducted. First of all, a baseline reference system *(BASE)* with sentence-by-sentence substraction of the mean values the cepstral coefficients. Then a temporal smoothing filter has been applied to it in order to see its separate effects *(BASE+TES)*. For the third experiment conducted, the cepstral coefficients have been equalized using HEQ and a Cumulative Density Function calculated over the average of the clean training data*(ECDF)*. 31 quantiles were estimated per utterance to be confronted with the ones of the reference CDF, defining in such a way a piecewise linear transformation for each coefficient.The last experiment *(AFE)* uses the ETSI standard advanced front-end parameterization algorithm[10]

3.2. Comparison between the possible filter locations

Tests have been done to compare the 3 possible scenarios on which to apply TES. The best results are obtained when performing the temporal smoothing once the features have been equalized to a clean data CDF (figure 2). The second best procedure is to smooth the features in the Gaussian domain (figure 1) and then equalize them again to clean CDF. The worst performance is obtained if the smoothing is done before the non linear transformation is applied(figure 3). This result makes sense as the equalized domain should ease the separation of the relevant and irrelevant information contained in the features. Equalization using a clean CDF prior to the filtering, produces better results than the equalization using a Gaussian reference. Table 1 shows comparative average WER for the 14 tests of AURORA4 clean training experiments (8KHz / 166 small tests) performed in the 3 possible scenarios: results for the equalization to a clean reference plus filtering (ECDF+TES), Gaussianization plus filtering plus equalization to a clean reference CDF (GAUSS+TES+ECDF), and filtering plus equalization to a clean reference (TES+ECDF) are exposed. As a consequence of this first analysis, the experiments conducted have used the processing flow *"HEQ plus TES"* depicted in figure 2

Table 1: Word error rates for the 3 possible combinations of equalization and temporal filtering.

ECDF+TES	GAUS+TES+ECDF	TES+ECDF
35,48	35,69	36,65

3.3. Filter sensitivity to noise

Figure 4 shows The impulse response of the temporal smoothing filter for different noise levels of a same sentence of AURORA2 database. For higher SNR the impulse response is closer to a delta-Dirac while when the SNR decreases the impulse response shape becomes smoother. The efficiency of the algorithm has been analyzed for different SNRs in order to see the benefits. In figure 5 we have plotted the relative WER reduction obtained from applying the temporal smoothing filter to the baseline features of AURORA2 at different SNRs. It can be seen that the optimal gain is obtained for a SNR of 5-10 dB. For higher SNRs, benefit decrease but still there is an improvement of 4% for 20 dB. For bad SNRs lower than 2,5 dB, the algorithm does not produce benefits. It increases the WER for very low SNRs.

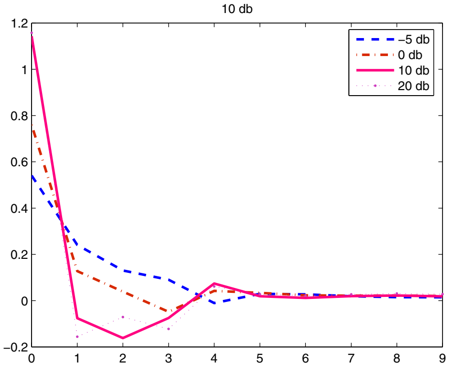

Figure 4: Impulse Response for SNR= -5dB, 0dB, 10dB and 20dB

Figure 6 shows the effect on the energy of the temporal filter applied to an AURORA4 sentence in a noisy environment (*Test07* in the picture) . The energies are depicted with and without HEQ processing. The clean sentence (*Test01* in the picture) is also depicted for comparison purposes.

3.4. Numerical results

Table 2 shows the WER obtained for the 14 test sets of AURORA4. A relative WER reduction of 4,6% is obtained when adding TES to the ECDF algorithm, making it closer to the AFE upper bound standard. If the TES normalization is applied directly over the baseline features, the improvement achieved is only slightly higher:4,9%.This similarity in the improvement points out that the temporal smoothing captures distortions that can not be eliminated with the equalization. The benefits of both normalizations can be added up in the case of AURORA4.

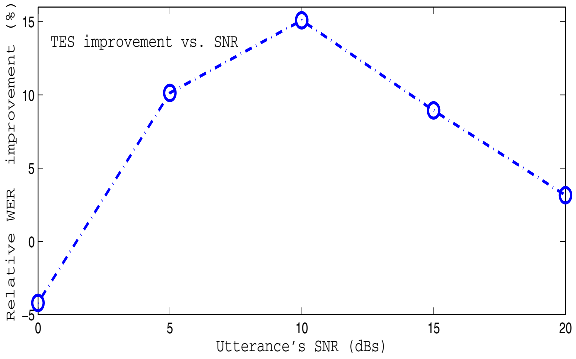

Figure 5: Improvement gained with TES at different SNRs

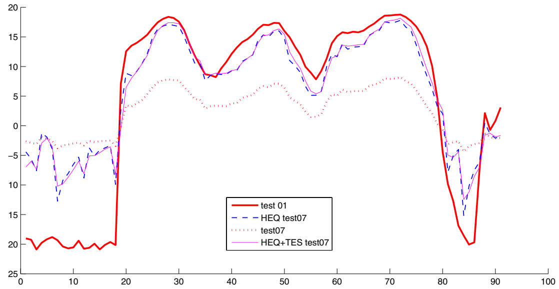

Figure 6: C_0 coefficient for the clean sentence and the 3 possible processing of a noisy sentence

The results for the 3 test sets of AURORA2 can be seen in Table 3. In this case the relative WER decreases a 12% when applying the temporal smoothing algorithm to the features already equalized via ECDF. If TES is applied to the baseline features, the improvement obtained is 6,5%. In this case, the benefits of both equalizations (TES and HEQ) are not orthogonal as it happened in AURORA4.

Table 3: Word error rates for the 3 sets of test of AURORA2 clean training experiment. Results for the baseline system (BASE),histogram equalization to a clean reference CDF (EDCF), the proposed temporal smoothing filter added to ECDF (ECDF+TES) and the ETSI advanced front-end (AFE).

	TEST A	TEST B	TEST C	Avg
BASE	36	30,9	35,27	33,82
BASE+TES	34	28,82	35,27	31,62
ECDF	17,06	17,3	18,97	17,54
ECDF+TES	16,24	14,21	16,35	15,45
AFE	12,49	12,94	14,48	13,07

Table 2: Word error rates for the 14 test of AURORA4 clean training experiment (8KHz / 166 small tests). Results for the baseline system (BASE), histogram equalization to a clean reference CDF(EDCF), the proposed temporal smoothing filter added to ECDF (ECDF+TES) and the ETSI advanced front-end (AFE).

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	Avg
BASE	13,22	24,68	46,00	47,62	52,67	44,8	54,73	22,58	36,21	55,40	58,31	65,34	54,11	62,28	45,57
BASE+TES	13,33	22,65	44,01	46,74	49,91	43,35	53,63	22,65	33,33	51,27	54,77	63,13	49,8	58,53	43,36
ECDF	11,82	22,62	37,75	38,90	36,91	37,46	40,92	21,29	32,67	45,93	49,28	50,61	44,60	49,65	37,17
ECDF + TES	11,42	21,40	35,25	37,53	34,59	36,17	38,56	20,15	28,80	43,87	47,66	49,83	44,75	46,77	35,48
AFE	12,7	17,8	30,4	34,8	30,6	34,9	31,7	18,8	25,1	38,0	44,9	40,4	39,3	38,4	31,3

4. Conclusions

This paper presents a new method to add and normalize temporal information contained in the voice features used in robust recognition. An study on the most convenient domain to perform this normalization is done, concluding that the histogram equalization transforms the MFCC features into a more operative domain where the distinction between noise and information is easier also when dealing with inter-frame correlations. The benefits obtained with this method are comparable with those of the few methods of normalization of the temporal information [8] found in the specific literature. A deeper analysis taking into account the coefficients temporal covariances instead of correlations, and different ways to define the normalization filter depending on the frame energy are undergoing for future works.

5. Acknowledgements

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-507943 (HI-WIRE, Human Input that Works in Real Environments) and SR3-VoIP projects (TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

6. References

[1] H. Hermansky, N. Morgan, A. Bayyman, and P. Kohn, "Rasta-plp speech analysis," *ICSI Technical Report*, , no. TR-91-069, 1991.

[2] S. van Vuuren and H. Hermansky, "Data driven design of rasta-like filters," in *Proc. of EUROSPEECH*, 1997, vol. 1, pp. 409–412.

[3] J.P. Openshaw, Z.P. Sun, and J.S. Mason, "A comparison of composite features under degraded speech in speaker recognition," *Acoustics, Speech, and Signal Processing,IEEE International Conference on*, vol. 2, pp. 371–374, 1993.

[4] C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Robust asr front-end using spectral-based and discriminant features: experiments on the aurora tasks," in *Proc. of EUROSPEECH*, 2001.

[5] A. de la Torre, J. Segura, C. Benitez, A. Peinado, and A. Rubio, "Non linear transformation of the feature space for robust speech recognition," in *Proc. of ICASSP*, 2002.

[6] A. de la Torre, A.M. Peinado, J.C. Segura, J.L. Perez-Cordoba, M.C. Benítez, and A. Rubio, "Histogram equalization for noise robust speech recognition," *IEEE Transactions on Acoustic, Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[7] S. Molau, M. Pitz, and H. Ney, "Histogram based normalization in the acoustic features space," in *Proc. of ASRU*, 2001.

[8] Y. Obuchi and R. Stern, "Normalization of time-derivative parameters using histogram equalization," in *Proc. of EUROSPEECH*, 2003.

[9] Y. Obuchi, "Improved histogram-based feature compensation for robust speech recognition and unsupervised adaptation," in *Proc. of ICSLP*, 2004.

[10] ETSI ES 202 050 Recommendation, "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm;," 2002.