MEDIAN FILTERING THE TEMPORAL PROBABILITY DISTRIBUTION IN HISTOGRAM MAPPING FOR ROBUST CONTINUOUS SPEECH RECOGNITION

Christian Arcos Gordillo, J R Boisson de Marca and Abraham Alcaim

Center for Telecommunications Studies of the Catholic University CETUC-PUC,
Rio de Janeiro, Brazil
Email: (christian, jrbdemarca, alcaim)@cetuc.puc-rio.br

ABSTRACT

The nonlinear distortion in the cepstral coefficients domain introduced by additive noise in the speech signal, results in high degradation performance in systems of Automatic Speech Recognition (ASR). For this reason, we propose a median filter which smooths the probability distribution functions of degraded features, thus reducing the mismatch between training data and test. The new proposal uses a histogram mapping to obtain the PDFs (probability distribution functions) of each feature vector and applies a nonlinear median filtering before mapping to the reference PDF. The algorithm efficiency is analyzed and compared to a recently proposed linear mean filtering technique on the PDFs. From the experimental results it can be concluded that the histogram smoothing through the median nonlinear filtering reduces the mismatch between training data and test, improving the system performance under adverse conditions.

Index Terms— noise-robust speech recognition, continuous speech recognition

1. INTRODUCTION

The Automatic Speech Recognition (ASR) has become in recent decades a research field of growing importance. The improvement of algorithms and more accurate models, compared with to the evolution of computing systems that are becoming more powerful and affordable, enable the development of more advanced man-machine interfaces. These interfaces allow advanced access to a great deal of information through a form of communication as natural as speech. It replaces some traditional interfaces based on keyboards, panels or similar devices, providing a great naturalness as well as a wide range of applications by several kinds of users in different operating environments.

The degree of complexity of ASR systems varies according to the number of words that are used and how sentences are uttered. In this field, one of the most critical application is Continuous Speech Recognition(CSR) [1], which has shown to have high levels of performance on clean conditions or low noise. However, one of the biggest problems that CSR systems present is the speech degradation when it is captured in highly noisy environments [2], making it more difficult to understand what is being said.

During the last decades a lot of approaches (e.g., cepstral nean normalization [3], spectral substraction [4], computacional auditory scene analysis [5], etc.) have been proposed trying to mitigate the effects of background noise. Those robust methods can be divided into three categories, Speech enhancement, Feature Compensation and Model Adaptation. Studies such as [6] show how compensation techniques of characteristics play an important role to minimize the statistical mismatch between the training data and tests.

Histogram mapping (HMAP) is a set of non-linear transformations of the probability distribution functions (PDF) for each of the features to a typical reference PDF function. Such transformation tend to compensate the distorsion the noise produced over the different components of the feature vector and improve the performance of the recognition system under noise conditions.

Recently in [7] the effect of a temporal linear mean filter applied to the PDFs of histogram equalization (HEQ) was studied. This technique improved the recognition system performance by reducing the mismatch caused by noise and preserving the important components of voice through reduced distortion at high modulation frequencies.

In this paper, our main focus is to improve the PDFs characterization of the systems mentioned above through a smoothing of components, using a non-linear filtering composed of a sliding window of a number of odd elements of the frame PDFs. This new approach called MED-HMAP replaces each current value of the probability distributions by the median of samples in the window. It was primarily motivated by the fact that the smoothing by median filtering is quite insensitive to high local noise intensities.

Traditionally, the Mel Frequency Cepstral coefficients (MFCC) have been used [8] to obtain the representation of the voice signal as a sequence of feature vectors that contains spectral information of short time periods and to apply the mapping techniques of probability distributions. However, its performance falls rapidly in the presence of noise. Recently,

The authors would like to thank CAPES, for the financed work.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

Kim in [9] introduced a more efficient method called Power-Normalized Cepstral Coefficients (PNCC) presenting them as an evolution from MFCC, changing some of the steps in the algoritm to make it more robust. These features represent in our work the feature vectors that will be modified to obtain a reference function.

This paper is organized as follows: In Section 2 we discuss prior work related to HEQ. The mean linear filtering on HEQ (FHEQ) is briefly revised. Section 3 details the proposed MED-HMAP technique. The experimental procedure and the discussion of results are given in Section 4 and finally, Section 5 draws some conclusions.

2. PRIOR WORK

2.1. Histogram Mapping

The philosophy behind HMAP is a nonlinear mapping of the features representing both characteristics of training and test voice, in order to adjust a common range. Due to degradation of the representation space caused by the additive noise the first order statistics is changed, generating mismatch between training data and test, and therefore lowering the ASR systems performance.

The HMAP is based on the technique known in image processing as Histogram Equalization (HEQ) [10]. Its theoretical fundamental is related to the properties of random variables in which a random variable *x* with probability density function (pdf) $p_x(X)$ and a PDF $C_x(X)$ can be transformed into a random variable $Y_{HEQ} = F(X)$ with a PDF of reference $C_y(Y)$. Defining a uniform random variable ω by the PDFs of *x* and *y*, i.e.

$$\omega = C_x(X) = C_y(F(X)) \tag{1}$$

we obtain the mapping equation [10]

$$Y_{HEQ} = F(X) = C_y^{-1}(C_x(X)) \tag{2}$$

2.2. Temporal mean filter in histogram equalization

Recently, the authors in [7] have proposed a method that integrates temporal linear filtering techniques with HEQ, called FHEQ. This proposal, motivated by the concepts of temporal filtering methods as RASTA [11] and ARMA [12], uses a low pass filter of the first order with coefficients 0.25 and 0.75 [7].

This new proposal has been successful by smoothing the original sequence of PDFs of each MFCC coefficient before mapping, reducing recognition errors since each point of the new PDF is a weighted sum of the two neighboring points of the original PDF. This proposal modifies equation (2) to

$$Y_{FHEQ_i} = C_y^{-1}(\sum_k h_k C_x(X_{i-k})) \tag{3}$$

where h_k represents the coefficients of the temporal filter, $C_x(X_i), i = 1, ..., N$ are the PDFs of an arbitrary feature sequence $X_1, ..., X_N$, N being the total number of frames (which is viewed as the sample set of a random variable x) and $Y_1, ..., Y_N$ is the new feature sequence with PDFs that approach the reference PDF.

3. MEDIAN FILTERING OF THE PDF IN HISTOGRAM MAPPING

HEQ is a robust technique used to transform the features into a reference domain less affected by changes in the acoustic environments, by normalizing the PDF of the each feature in such a way that the acoustics environment affect are removed. However, there is a distortion in the adjustment between the original PDF and the reference PDF.

In a recent work [7], a technique was presented that can greatly reduce the distortion caused by mapping. This procedure involves temporal filters. As seen in previous section, a temporal mean low-pass filter of two points was used in order to obtain a new softened sequence of the original PDFs. However, in these filters every point of the probability function that was degraded may significantly vary. Consequently, the mean can also greatly differ from the PDF values. This means that this type of filter is quite sensitive to local changes.

In our work, the respective PDFs of test data are smoothed through a median filter. Each point of the reference PDF is generated by calculating the median of the values of PDFs of the neighborhood around the corresponding point of the original PDF. Median filtering is expected to provide a better smoothing than mean filtering in very noisy environments. This is because smoothing in median filtering is less sensitive to high local noise intensities.

The median filter on the PDF is obtained for a sliding window with an odd number of samples through the signal. It replaces the middle sample, by the median of the samples in the window. Using this operation and equation (2) to take into account the histogram mapping, we obtain

$$Y_{MED-HMAP_i} = C_y^{-1}(Med[C_x(X_{i\pm k})], k \in W) \tag{4}$$

where *W* is the window around the PDF of each feature coefficient and *Med* is the median function. The median filtering algorithm has to sort the sample values of the PDF of the window in ascending order, and take the intermediate value. In this work we select $k = 3$.

4. EXPERIMENTS

4.1. Databases

The performance of the MED-HMAP method is evaluated through recognition experiments under noise conditions using two databases.

The TIMIT [14] corpus was created especially for experiments on continuous speech recognition independent of speaker. It has a total of 6300 sentences spoken by 630 people of which 70% are men and 30% women, where everyone speaks 10 phrases covering the different accents of American English for both sexes. For this study the database was divided into training set and testing set. In 4620 sentences speech was used to train and to create the language model. The testing data of 1680 sentences was corrupted by noise of different environments.

This work selects different background noises from NOISEX-92 corpus, which contains sound files of various kinds of noise, like white, babble, f16, and factory. To get corrupted samples test, we just take the samples of the clean voice and add a noise signal on it.

4.2. Experimental setups

4.2.1. Front-End

The front-end is based on PNCC parametrization where the voice signal was sampled at 8 KHz and segmented into frames that are represented by a feature vector with the following parameters:

- PNCC features use a frequency domain 30-bands gammatone filter bank that analyzes the speech signal at each 10ms with a 25.6ms time window.
- Short term spectral powers were estimated by integrating the squared gammatone responses, and the resultant was compressed using 1/15 root. Only the first 20 values of the discrete cosine transform (DCT) have been considered.
- Finally, the delta features were included, doubling the number of values per frame.

4.2.2. CSR system setup

The Continuous Speech Recognition System was implemented with the HMM tool kit(HTK) [15]. For experiments, Hidden Markov Models (HMM) were built with three states from left to right using eight Gaussian Mixture Models (GMM) in order to represent initially monophones models and from these models to estimate the triphones of the English language with intra-word settings.

A trigram language model was estimated from the 4620 training phrases of the TIMIT data base.

Finally, the reference probability distribution chosen for mapping histograms was the Gaussian distribution with zero mean and unit variance, since this has a major advantage due to the fact that in most systems the output distributions of the hidden Markov model (HMM) are modeled as a Gaussian mixture.

We have used the classical GMM-HMM system to evaluate the proposed feature compensation technique. However further research is being conducted using the more recent DNN-based scheme.

4.3. Results and Discussion

The recognizer performance was assessed by the word accuracy rate (WAR), which is the proportion of correct words in test sentences. It is given by

$$WAR(\%) = 100\frac{N-(S+D+I)}{N} \tag{5}$$

where *N* is the number of words in the test, *S* is the number of substituted words , *D* is the number of deleted words and *I* is the number of inserted words.

In order to apply the compensation methods described in this paper, four types of tests were performed. In the first case, a reference system (*baseline*) based solely on PNCC features is used in clean conditions and subsequently corrupted with different kinds of noise at four different SNRs (0, 5, 10 and 15dB). The *baseline* results are obtained without any robustness technique, this means, the acoustic models are trained and tested on the above mentioned PNCC features. The other tests were carried out in order to compare the performance of our algorithm with the recently proposed in [7]. The same system conditions were used, but applying the HEQ, FHEQ [7] and the proposed MED-HMAP methods.

The data in Table 1 show recognition rates (WAR%) of the compensation methods based on histogram mapping, for each type of noise averaged over 0, 5, 10 and 15dB.

Table 1. Recognition results obtained for the TIMIT database. Avaraged over the different conditions of SNR

	white	**babble**	**f16**	**factory**
Baseline	59,64	66,52	70,02	58,61
HEQ	62,88	69,22	73,63	62,08
FHEQ	65,98	69,82	74,66	60,8
MED-HMAP	**66,5**	**70,42**	**74,73**	**61,88**

From Table 1 we can see that the techniques based on histogram mapping are more robust under all scenarios. Techniques using filters on PDFs before mapping give better results than the traditional HEQ technique.

On the other hand, the results show the importance of smoothing PDFs, before applying respective mapping. The proposed MED-HMAP system provides the bests results in performance. It increases recognition average from 62.88% to 66.50% in white noise, from 69.22% to 70.42% in babble noise, from 73.63% to 74.73% in f16 noise, and for noise factory shows no improvement over the traditional HEQ.

Table 2 shows the behavior of the recognition system in function of the SNR level. The results are averaged over all the different kinds of noise.

Table 2. Recognition results obtained for the TIMIT database. Avaraged over the different kinds of noise.

SNR	Baseline	HEQ	FHEQ	MED-HMAP
0	31,31	35,89	41,09	42,35
5	63,19	66,06	66,86	68,58
10	78,30	81,59	80,04	80,64
15	81,99	84,27	83,19	82,43
clean	86,58	89,18	88,28	87,37

From Table 2 it can be seen that the proposed MED-HMAP technique overperforms the HEQ and FHEQ schemes for low SNR (less than 10dB). When filtering methods are applied to the clean signal or where the degradation of voice is not severe, important features are removed from the signal, thus causing loss of speech recognition performance. For this reason at higher SNR the other schemes overperform the MED-HMAP technique.

5. CONCLUSIONS

We have proposed a feature compensation method based on a nonlinear smoothing of the probability distribution functions in the technique of histogram mapping. This new proposal, a median filtering of the PDFs (MED-HMAP), was shown to be a promising technique. It significantly improves the recognition system, being able to increase success rates of speech by an average of 5% with respect to a *baseline* system. It also yields a superior performance at low SNR (less than 10dB) when compared to the HEQ technique and the recently proposed [7] mean filtering scheme. Finally, it is also important to remark that the MED-HMAP algorithm is quite simple in architecture.

REFERENCES

[1] R. Cole, J. Mariani, H. Uszkoreit, G. Varile, A. Zaenen, and A. Zampolli, *Survey of the state of the art in human language technology*, Cambridge University Press, vol. 12. 1998.

[2] M. Vikramjit, H. Franco, M. Graciarena, and D. Vergyri, *Medium-duration modulation cepstral feature for robust speech recognition*, in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1749–1753, 2014.

[3] F.H. Liu, A. Acero, and R. Stern, *Efficient Joint Compensation of Speech For the Effects of Additive Noise and Linear Filtering*, in Proc. International Conference on Acoustics, Speech and Signal Processing, (ICASSP), Vol. 1. 1992.

[4] N. Virag, *Single channel speech enhancement based on masking properties of the human auditory system*, Speech and Audio Processing, IEEE Transactions, pp 126-137, 1999.

[5] S. Soundararajan, and D. Wang. *Transforming binary uncertainties for robust speech recognition*, Audio, Speech, and Language Processing, IEEE Transactions, pp 2130-2140 2007.

[6] J. Segura, C. Benítez, A. de la Torre, A. Rubio, and J. Ramírez, *Cepstral domain segmental nonlinear feature transformations for robust speech recognition*, Signal Processing Letters, IEEE, vol. 11, no. 5, pp 517–520, 2004.

[7] S. Wang, Y. Tsao, J. Hung, *Filtering on the temporal probability sequence in histogram equalization for robust speech recognition*, in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 1749–1753, 2013.

[8] W. Han, C. Chan, C. Choy, and K. Pun *An efficient mfcc extraction method in speech recognition*, in International Symposium on Circuits and Systems (ISCAS), pp 4 – pp, 2006.

[9] W. Kim, and R. Stern *Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribuition and on power flooring*, in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010.

[10] A.K. Jain *Fundamentals of digital image processing*, Prentice-Hall Englewood Cliffs, vol. 3, 1989.

[11] L. Hermansky, and N. Morgan *Rasta processing of speech*, IEEE Transaction on Speech and Audio Processing, vol. 2, pp 587–589, 1994.

[12] C. Chen, K. Filali, and J. Bilmes *Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases*,in Proc. International Conference on Spoken Language Processing (ICSLP), pp. 241 – 244, 2002.

[13] Md. Molla, M. Pitz, and H. Ney *Histogram based normalization in the acustic feature space*,in Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, 2001.

[14] G. John, et al. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT) Training and Test Data*, 1993.

[15] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland *The htk book*, Cambridge University Enginnering Department, vol. 3, 2002.