INTERSPEECH 2009 BRIGHTON

Progressive Memory-Based Parametric Non-Linear Feature Equalization

Luz Garcia[2], *Roberto Gemello*[1], *Franco Mana*[1], *Jose Carlos Segura*[2]

[1]LOQUENDO, Torino, Italy.
roberto.gemello@loquendo.com, franco.mana@loquendo.com

[2]Department of TSTC, University of Granada, Granada, Spain.
luzgm@ugr.es, segura@ugr.es

ABSTRACT

This paper analyzes the benefits and drawbacks of PEQ (Parametric Non-linear Equalization), a features normalization technique based on the parametric equalization of the MFCC parameters to match a reference probability distribution. Two limitations have been outlined: the distortion intrinsic to the normalization process and the lack of accuracy in estimating normalization statistics on short sentences. Two evolutions of PEQ are presented as solutions to the limitations encountered. The effects of the proposed evolutions are evaluated on three speech corpora, namely WSJ0, AURORA-3 and HIWIRE cockpit databases, with different mismatch conditions given by convolutional and/or additive noise and non-native speakers. The obtained results show that the encountered limitations can be overcome by the newly introduced techniques.

Index Terms— Histogram Equalization, Parametric Equalization , Feature Normalization, Robust Speech Recognition.

1. INTRODUCTION

PEQ (*Parametric Non-linear Equalization*) was first introduced in [5] with the aim of improving the results of HEQ (*Histogram Equalization*) [2], as normalization algorithm for the MFCC features. HEQ is very effective if the amount of speech material to be normalized is sufficient to perform a robust estimate of the histogram. If only a smaller amount of speech is available, the parametric expression of the pdfs used by PEQ provides more trustable statistics. Nevertheless some drawbacks of PEQ were left to solve: for some databases, even the parametric expression of the probability density functions was not trustable as it had to be calculated with very short or noisy utterances. On the other hand, for clean conditions both HEQ and PEQ (although occurring to a lower extent in the case of PEQ), introduce a non desired distortion intrinsic to the normalization process. The purpose of this work is to present some improvements of the original PEQ that successfully overcome the drawbacks mentioned. The problem of unreliable statistics of very short sentences is approached by Memory PEQ which consists of a recursive estimate of global statistics over the sequence of utterances, and a mixed use of that global statistics together with the local sentence statistics to perform PEQ equalization. Progressive PEQ, instead, is introduced in order to limit the intrinsic distortion of the normalization process. That is accomplished by considering for PEQ equalization only the features whose variance is greater than the variance of the distortion introduced by the normalization process itself.
The work is organized as follows: section 2 briefly describes the philosophy of PEQ. Section 3 proposes two evolutions of the original algorithm and their benefits. Section 4 presents some experiments and analyzes their results. Finally conclusions on the suitability of the approach are presented in Section 5.

2. PARAMETRIC EQUALIZATION

PEQ reduces the mismatch between training and test conditions by transforming the statistics of each test utterance (*local statistics*) in order to match the statistics of the training set (*reference statistics*). The peculiarity of PEQ consists of assuming a bimodal Gaussian distribution for the probability density functions of the MFCC parameters. The reference statistics are therefore composed by the mean and variance of the Gaussian describing the silence frames ($\mu_{n,x}$ and $\sum_{n,x}$) and mean and variance of the Gaussian describing the voice frames ($\mu_{s,x}$ and $\sum_{s,x}$). The local statistics of the utterance to be normalized are defined as well with two Gaussian representing the silence frames ($\mu_{n,y}$ and $\sum_{n,y}$) and the voice frames ($\mu_{s,y}$ and $\sum_{s,y}$).

The novelty of this technique lies in the method used to classify the frames as speech or non-speech. The usage of a Voice Activity Detector has shown that using hard decision to classify the frames produces a discontinuity around the speech frames with a high noise level, leading to poor recognition rates. Bo Liu used in [3] the EM algorithm to calculate independent class probabilities for each Cepstral Coefficient with results similiar to those of HEQ. PEQ proposal is to

utilize the Cepstral coefficient *C0*, which captures the frame logaritmic energy, to catalogue frames as silence or speech frames. The posterior probabilities $P(n|y)$ and $P(s|y)$ are obtained using a simple two-class Gaussian classifier on the *C0* term. After initializing the silence and noise classes with frames below and above the *C0* average, EM re-estimation is iterated until convergence. The linear transformation produced by PEQ on a test vector y originates a normalized vector $\hat{x}$ with the following expression in case y is a silence frame:

$$\hat{x_n} = \mu_{n,x} + (y - \mu_{n,y})(\frac{\sum_{n,x}}{\sum_{n,y}})^{\frac{1}{2}} \quad (1)$$

For the case of y being a voice frame, the expression of the normalized vector is:

$$\hat{x_s} = \mu_{s,x} + (y - \mu_{s,y})(\frac{\sum_{s,x}}{\sum_{s,y}})^{\frac{1}{2}} \quad (2)$$

The normalized frame $\hat{x}$ will be the weighted average considering both probabilities of the frame being silence or voice:

$$\hat{x} = P(n|y) \cdot \hat{x_n} + P(s|y) \cdot \hat{x_s} \quad (3)$$

3. PARAMETRIC EQUALIZATION EVOLUTIONS

3.1. Progressive PEQ

Despite the benefits of the equalization process, some additional distortion is introduced when equalizing features. There are two factors responsible for this additional distortion:

i) The reference CDF is assumed to be a bimodal Gaussian distribution. This is a parametric approximation of the real CDF.

ii) The linear transformation defined to equalize the MFCC parameters is calculated using the parametric approximation of the reference CDF and the parametric approximation of the test utterance CDF. To the distortion introduced by using parametric approximations, we must add the fact that the test utterance statistics are calculated with the few data contained in a single utterance.

In order to minimize this distortion, the possibility of equalizing only certain coefficients has been studied, rendering good results. The deep analysis of the MFCC coefficients concludes that they have different statistical properties and a variable discriminative capacity, being affected in a dissimilar manner by noise. The lower order cepstral coefficients have bigger variance and higher discriminative power than the higher order coefficients. Coefficient *C0* represents the energy of the signal, while *C1* represents the global energy balance between high and low frequencies. The rest of *Cis* are not easy to identify with concrete aspects of voice production or perception. They keep spectral details which permit the distinction between similar sounds. This analysis concludes that when the variance of the distortion introduced by PEQ is similar to the variance of the feature, its normalization is useless. According to that, only the features conveying more information have been normalized in this work, while the rest of features were left unchanged. Figure 1 shows the evolution of the word error rate reduction obtained in the recognition experiments with HIWIRE database [4] when equalizing only a certain number of Cepstral coefficientes. The best results have been obtained normalizing only the Energy (or *C0*) and the first four MFCCs. This partial normalization has been defined as *Progressive PEQ*.

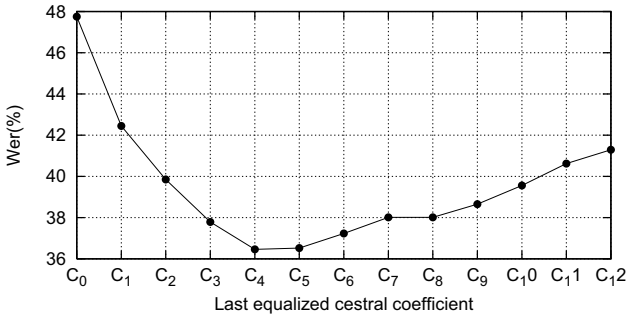

Fig. 1. WER for progressive PEQ for HIWIRE database

3.2. Memory PEQ

Besides the distortion intrinsic to the normalization process, the main limitation of the standard PEQ is the poor accuracy of the local statistics provided by the test utterance. One of the basic assumptions of the parametric normalization is that the statistics computed on the speech features are independent on what was actually spoken, (i.e. the ratio of phoneme observations is similar in training and test). To approximate this condition, a certain amount of voice is needed (1 minute would be good, 10-20 seconds are an acceptable approximation).

The problem is that we often have only few seconds of speech in the local utterance, sometimes a single word. In those cases, in order to make statistics more accurate, one possible solution is to capture the test utterance's evolution in a longer time term and use that evolution in conjunction with the local statistics to normalize the current utterance with PEQ. We name this method of computing statistics *Memory PEQ (MPEQ)*. This can be accomplished with the following scheme:

i) *Local(t)* represents the *local statistics* computed using only the utterance at time *t*.

ii) *Memory(t)* stands for the *global statistics* computed taking into account the statistics computed on the past

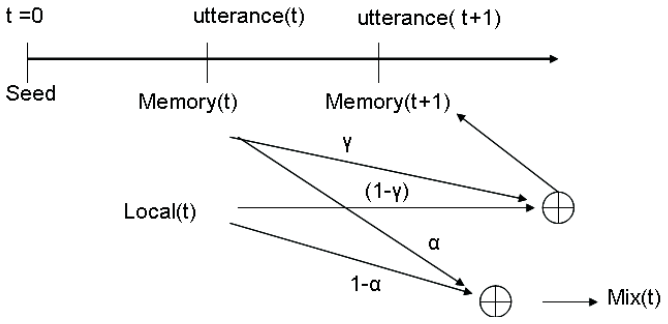

Fig. 2. Flow for Memory PEQ

utterances.

$$Memory(t+1) = \gamma \cdot Memory(t) + (1-\gamma) \cdot Local(t)$$

where γ determines the dynamicity of Memory. The Memory statistics are the global statistics that evolve in time and become utterance by utterance more accurate.

iii *Seed* statistics used to initialize the Memory. Memory(0)= Seed. Usually seed is set to training set statistics.

iv) *Mix(t)* are the *balanced local statistics* computed as a mixture of the *global* and *local statistics* according to the following rule:

$$Mix(t) = \alpha \cdot Memory(t) + (1-\alpha) \cdot Local(t)$$

where α determines the balance between Memory and Local statistics. Mix(t) statistics are used to normalize the local utterance in MPEQ, instead of the Local(t) statistics of standard PEQ.

This scheme allows to clearly separate the estimation of the global statistics (*Memory(t)*) and its usage, with two different parameters: γ determines the dynamicity of *global statistics* evolution, while α determines how to balance the use of the global statistics *Memory(t)* and the use of the *local statistics* Local(t) to produce the *balanced local statistics* Mix(t).

4. EXPERIMENTS AND RESULTS

Experimental activity has been devoted to test PEQ normalization and the proposed evolutions. The recognition system used is the state-of-the-art Loquendo ASR system [1] which uses acoustic models based on a hybrid combination of Hidden Markov Models and Multi Layer Perceptron. Phonetic units are stationary-transitional units made up by phonemes plus diphone transitions between them. When used, PEQ normalization is applied both in training and in test.

4.1. Test corpora

The following test corpora have been employed in the experiments:
WSJ0 5K:
Train: Standard WSJ0 SI_84 train set, Sehnheiser microphone, 7236 sentences.
Test: SI_ET_05 test set, 8 speakers, and 40 sentences per speaker. Two channels: WV1, Senheiser microphone (matched condition) and WV2, other microphones (mismatched condition). Vocabulary: 5K words, with standard trigram LM from Lincoln labs.
AURORA3 8Khz Italian: Connected digits in car environment. Signal collected by hand free (ch1) and close talk (ch0) microphones, downsampled to telephonic band.
Train: Telephonic corpora not related to AURORA3 (SpeechDat).
Test: Standard "well-matched" list is used, divided into ch0 (664 utterances) and ch1 recordings (645 utterances). The mismatch condition present on this test is additive (car) noise that affects ch1 subset.
HIWIRE cockpit database[4]: Noisy and non native English speech corpus for cockpit communications. It includes short vocal sentences in English, corresponding to aeronautic commands. 81 non-native speakers from 4 countries.
Train: as HIWIRE does not have a training component, WSJ0 and TIMIT training sets have been used.
Test: Four noise conditions are tested: Clean, Low Noise (SNR=10dB), Medium Noise (SNR=5dB) and High Noise (SNR=-5 dB). The test set has 4049 utterances for each condition. The mismatch condition present in this test set is additive (aircraft) noise. An additional problem is the presence of short sentences that makes difficult a reliable estimation of feature statistics for normalization purposes.

4.2. Test conditions

The following normalization conditions have been used in the experiments and are reported in the result tables:

- ◇ NO PEQ:(*NO_PEQ*): no normalization is applied.
- ◇ STANDARD PEQ:(*PEQ_STD*): baseline PEQ normalization [5] is applied.
- ◇ PROGRESSIVE PEQ:(*PEQ_EC4*):PEQ that normalizes only *C0* and the first 4 cepstral coefficients.
- ◇ MEMORY + PROGRESSIVE PEQ:(*MPEQ_EC4_05_09*): Memory built on progressive PEQ is applied. The seed of the memory are statistics computed on the training set. The memory has a high inertia ($\gamma = 0.9$). The statistics used to normalize are a 50% mix of Memory and Local Statistics: Mix(t)=α*Memory(t)+(1-α)*Local(t), with α=0.5.

WSJ0 5k - trigram LM			
Normalization	**WV1**	**WV2**	**AVG**
NO_PEQ	93.0	70.3	81.6
PEQ_STD	92.5	78.0	85.2
PEQ_E4C	93.2	77.9	85.5
MPEQ_E4C_05_09	93.0	77.9	85.4

Table 1. Word Accuracy results for WSJ0

AURORA3 Well Matched Test Set			
Normalization	**CH0**	**CH1**	**AVG**
NO_PEQ	98.5	84.5	91.5
PEQ_STD	99.5	88.1	93.8
PEQ_E4C	99.3	90.1	94.7
MPEQ_E4C_05_09	99.4	90.5	94.9

Table 2. Word Accuracy results for AURORA3 Italian

4.3. Discussion

Table 1 shows the normalization results for WSJ0 database for which the following behaviors are observed: for Matched Conditions (WV1), Standard PEQ decreases its performance in terms of error reduction (*E.R*) (-7.2% *E.R.*) due to the distortion introduced by the normalization. Nevertheless, Progressive PEQ maintains the performance (+2.8% E.R.). For Mismatched Conditions (WV2), Standard PEQ increases the performance (+25.9 % E.R.) very similarly to Progressive PEQ (+25.6% E.R.). For this database, the use of Memory PEQ is not very important as the test utterances are long enough to correctly estimate local statistics.
Table 2 shows the results on AURORA3 Italian 8Khz database. In this case, the acoustic models have been trained with telephonic corpora not related to AURORA3 (SpeechDat Italian) and PEQ normalization is always beneficial as there is a clear mismatch between training and test conditions. Highest improvements are obtained on the noisy channel (ch1). MPEQ_E4C_05_09 produces the best results (38.7% E.R.).
Table 3 shows the results for HIWIRE database. For this database PEQ evolutions achieve important benefits. For clean tests Standard PEQ works slightly worse, due to the distortion introduced, but Progressive PEQ (PEQ_E4C) maintains performances. On noisy data PEQ_STD generates a 11.3% error reduction, PEQ_E4C obtains an 18.5% error reduction, and MPEQ_E4C_05_09 accomplishes an error reduction of 23.0%. The benefit of PEQ evolutions is more evident for this database because it is characterized by the presence of many short phrases or even single words. In this case the usage of Memory PEQ is important as the local statistics are not completely reliable and too much dependent on the phonetic contents of the utterance.

HIWIRE Cockpit non-native database					
Norm.	**Clean**	**Low**	**Mead**	**High**	**AVG**
NO_PEQ	89.2	69.6	53.7	15.4	57.0
PEQ_STD	85.2	73.7	59.5	19.8	59.5
PEQ_E4C	87.7	77.0	63.9	23.3	63.0
MPEQ_E4C_05_09	89.6	78.9	66.4	24.9	64.9

Table 3. Word Accuracy results for HIWIRE

5. CONCLUSIONS

PEQ is a normalization method that already proved to be very efficient in equalizing in a blind way the differences between the local utterance and the training set conditions [5]. This paper faced two problems still open: the intrinsic distortion introduced by PEQ and the lack of reliability of local statistics in case of short sentences, where the vocal material is insufficient and the phoneme frequencies are not similar to those of the training. Two evolutions have been proposed to deal with this problem, namely progressive PEQ and Memory PEQ. Results on selected test cases presenting mismatch problems due to convolutive and additive noise as well as non-native speakers show that progressive PEQ always outperforms standard PEQ, while Memory PEQ obtains the best performances in case of short test sentences.

6. ACKNOWLEDGEMENTS

This research work was partially supported by the European Union under the IST-EU STREP program 'HIWIRE', contract number IS-2002-507943.

7. REFERENCES

[1] D. Albessano, R. Gemello, and F. Mana. Hybrid hmm-nn modelling of stationary-transational units for continuous speech recognition. *Int. Conference On Neural Information Processing*, pages 1112–1115, 1997.

[2] Angel de la Torre et al. Histogram equalization for noise robust large vocabulary speech recognition. *IEEE Trans. On Speech and Audio Processing*, 2003.

[3] Bo Liu et al. Double gaussian based feature normalization for robust speech recognition. *Proceedings of ISCLP'04*, pages 253–246, 2004.

[4] J.C. Segura et al. The hiwire database, a noisy and non-native english speech corpus for cockpit communications. *online at http://www.hiwire.org/.*

[5] L. Garcia and J. Segura et al. Parametric nonlinear feature equalization for robust speech recognition. *Proceedings of ICASSP'06*, pages 529–532.