ON-LINE MEMORY-BASED PARAMETRIC EQUALIZATION TO MULTIMODAL TRAINING CONDITIONS

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

[1]Loquendo S.pA., 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 describes the conceptual and algorithmic evolutions of Memory Based Parametric Equalization (MPEQ) needed to exploit the potentialities of the method within the state-of-the-art Loquendo ASR. MPEQ is the memory-based evolution of Parametric Non-Linear Equalization (PEQ) introduced to overcome the problem of unreliable statistics estimation in presence of very limited acoustic information in the test utterance to be normalized. The main limitations of the method that prevented its practical application were the lack of online implementation, the unrealistic unimodal assumption about the training statistics, the unconditioned application of equalization, and the need for retraining the acoustic models.

The paper describes how these limitations have been overcome and reports a large experimentation on many corpora that shows improvements in a variety of mismatched conditions, while preserving performances in matched conditions.

1. INTRODUCTION

The intrinsic variability of speech is faced in state-of-the-art commercial ASR systems mainly with a data intensive approach, employing huge multi-style training sets to estimate the statistical acoustic models.

A second more algorithmic approach is the use of some normalization method to reduce the distance between the test conditions and the training conditions [2].

Among these methods, Parametric Non-Linear Equalization (PEQ) was first introduced in [3] with the aim of improving the results of Histogram Equalization (HEQ) [1] as normalization algorithm for spectral derived features, like MFCC or RPLP. HEQ is very effective if the amount of speech material to be normalized is sufficient for a robust estimate of the histograms. Otherwise, the parametric Gaussian assumption of feature distributions at the bases of PEQ can provide more trustable statistics. However, if the acoustic material amounts only to a few seconds, also the parametric assumption can be insufficient for a robust estimate of test data feature distribution. The memory based evolution of PEQ, namely MPEQ, was introduced in [4] to overcome this problem, relying on the assumption of stationarity or slow variability of test conditions. Experimental evidence demonstrated the potentialities of MPEQ. Nevertheless, some drawbacks were still present to prevent the use of the method within a commercial ASR system. First MPEQ (like PEQ and HEQ) had no online implementability, i.e. it was not suited for real-time processing of audio input, but requested the prior acquisition of the entire test utterance before performing the normalization. Second, normalization was always performed, also in the case of well-matching conditions between test and training data, introducing in that case some decrease of performances, due to the intrinsic distortion introduced by every normalization. Third, there was an implicit assumption of unimodality of the training data distribution that is simply not true in the huge multi-style training sets. Last, it required the re-training of ASR models applying the normalization algorithm also in training. This aspect is usually not considered, but has a dramatic impact on a commercial ASR production process. The purpose of this work is to overcome these drawbacks and propose algorithmic evolutions that make PEQ practically usable inside a high performance real-time ASR system.

2. PARAMETRIC EQUALIZATION

2.1 Standard PEQ

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*) [3]. 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 of the mean and standard deviation of the Gaussian describing the silence frames ($\mu_{n,x}$ and $\sigma_{n,x}$) and mean and standard deviation of the Gaussian describing the voice frames ($\mu_{s,x}$ and $\sigma_{s,x}$). The *local statistics* of the utterance to be normalized are defined as well with two Gaussian representing the silence frames ($\mu_{n,y}$ and $\sigma_{n,y}$) and the voice frames ($\mu_{s,y}$ and $\sigma_{s,y}$).

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:

$$\hat{x}_n = \mu_{n,x} + \left(y - \mu_{n,y}\right)\frac{\sigma_{n,x}}{\sigma_{n,y}} \tag{1}$$

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

$$\hat{x}_s = \mu_{s,x} + \left(y - \mu_{s,y}\right)\frac{\sigma_{s,x}}{\sigma_{s,y}} \tag{2}$$

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

$$\hat{x} = P\left(n \mid y\right)\cdot\hat{x}_n + P\left(s \mid y\right)\cdot\hat{x}_s \tag{3}$$

The posterior probabilities P(n|y) and P(s|y) are obtained in standard PEQ using a simple two-class Gaussian classifier on the C_0 cepstral coefficient. After initializing the silence and voice classes with frames below and above the C_0 average, EM re-estimation is iterated until convergence.

2.2 Memory Based PEQ

The main limitation of standard PEQ is the poor accuracy of the

local statistics provided by the test utterance. The problem is that often only few seconds of voice are available in the test utterance, sometimes even a single word. This makes impossible to compute accurate *local statistics*. A solution to this problem was proposed in [4] with the use of Memory PEQ (MPEQ). The method keeps track of the evolution of the local statistics with an iterative average across the past utterances. The memory term, named *global statistics (gs*) is computed as a recursive linear combination with the *local statistics (ls)* according to the following formula:

$$gs(t+1) = \gamma \cdot gs(t) + (1-\gamma) \cdot ls(t) \qquad (4)$$

where γ determines the dynamicity of *gs*. Typically $\gamma = 0.9$.

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

$$bls(t) = \alpha \cdot gs(t) + (1-\alpha) \cdot ls(t) \qquad (5)$$

where α determines the balance between the memory term and test utterance statistics. Finally *bls* is used in MPEQ to normalize the test utterance, instead of the *ls* used by standard PEQ. Results reported in [4] show significant improvements in presence of short test sentences that make difficult a reliable estimation of the utterance statistics.

3. NEW EVOLUTIONS

Starting from the achievements presented in [4], we introduce in this work some new improvements of PEQ devoted to:

1. apply it in real-time;
2. deal with the multimodal training conditions of large training sets;
3. limit PEQ application only when necessary.

3.1 On-line MPEQ

The standard implementation of PEQ can only be applied off-line, i.e. the whole utterance must be acquired before applying the normalization. The proposed evolution makes PEQ suitable for on-line real-time application.

In standard PEQ there are two steps that are intrinsically performed off-line:

1) The computation of *local statistics (ls)*, that needs to analyze the whole sentence;
2) The probabilistic C_0 based VAD that is estimated with an iterative method that looks at all the utterance frames.

On the contrary, the PEQ normalization itself (formula (3)) is suited for on-line processing, as it can be applied frame-by-frame left-to-right without delays.

To produce an on-line implementation of PEQ, points 1) and 2) must be changed.

Two solutions could be devised to overcome point 1):

- The first one is to delay recognition a certain time window sufficient to estimate *ls* on that window (at least 1 sec from the start of speech should be needed). We discarded this solution because we cannot accept such a delay.
- The second one is to implement PEQ normalization not using the test utterance *local statistics (ls),* but statistics computed on the previous utterances, i.e. *global statistics* (*gs)* as defined in equation (4). That can be done under the hypothesis of stability (or slow variability) of the speech conditions (channel, noise, speaker). This hypothesis is true in many applicative scenarios, like automotive interactions, air traffic control, etc. If the conditions change, such alteration should be detected and a reset of *gs* should take place.

Point 2) is easier to be faced, as many ways exist to implement probabilistic VAD on-line. In this work we employ a neural network VAD, i.e. a recurrent MLP specifically designed and trained for voice presence probability estimation, as described in [5]. MLP VAD performances compares favorably with energy based VADs.

3.2 Equalization to Multimodal Training Conditions

PEQ normalization aims to reduce the mismatch between training and test conditions. Training conditions are usually represented by the reference statistics quadruple $rs = (\mu_{n,x}, \mu_{s,x}, \sigma_{n,x}, \sigma_{s,x})$. But this formulation assumes that the training set statistics are unimodal and can be represented by two multi-variate Gaussians with diagonal covariance matrix (one for noise and one for voice). This assumption is acceptable for small corpora, but it is not true for state-of-the-art speech recognition systems that are characterized by huge multi-style training sets, composed by several different components, recorded in different conditions (different transmission channels: PSTN, GSM, VoIP; different environments: home, office, automotive, etc; different microphones and devices).

Thus we extend the *reference statistics* concept formalizing it with a set of quadruples, one for each different training set component $RS=\{rs_1, rs_2, \ldots, rs_k\}$.

The idea is to normalize the test utterance towards the nearest $rs_j \in RS$ in order to reduce the normalization error. The distance between *gs* and the *RS* components is computed with a distance defined between probability distributions.

Let us assume $S_1 = ((\mu_{n,1},\sigma_{n,1}), (\mu_{s,1}, \sigma_{s,1}))$ and $S_2 = ((\mu_{n,2},\sigma_{n,2}), (\mu_{s,2}, \sigma_{s,2}))$. We can define the distance between them as:

$Dist(S_1, S_2) = \xi \cdot D((\mu_{n,1},\sigma_{n,1}), (\mu_{n,2},\sigma_{n,2})) + (1-\xi) \cdot D((\mu_{s,1},\sigma_{s,1}), (\mu_{s,2},\sigma_{s,2}))$, where ξ balances the noise and voice component of the distance.

Some suitable distances are the following:

1. Mahalanobis distance

$$D((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \sqrt{\frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2)}}$$

2. Bhattacharyya distance

$$D((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \frac{1}{4}\frac{(\mu_1-\mu_2)^2}{(\sigma_1^2+\sigma_2^2)} + \frac{1}{2}\ln\left(\frac{(\sigma_1^2+\sigma_2^2)/2}{\sigma_1\sigma_2}\right)$$

3. Kullback-Leibler distance (KLD)

$$D((\mu_1,\sigma_1),(\mu_2,\sigma_2)) = \frac{1}{2}\left[\frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} - 2 + (\mu_1-\mu_2)^2\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)\right]$$

3.3 Distance activated equalization

Provided that normalization always introduces an amount of distortion, it is advisable to apply it only if the distance between the *global statistics* (*gs)* and the nearest component of the *reference statistics RS* exceeds a given threshold. Otherwise it is better not applying the normalization, as the test utterance condition was already well represented in the training set.

3.4 Context Switch Detection

If a change in channel, noise condition or speaker happens (context switch) it has to be detected to avoid using *global statistics* no more coherent with the new conditions. The context

switch detection is performed monitoring the distance between *gs* and a short term average of *local statistics (lsa)*.

$$lsa(t+1) = \rho \cdot gs(t) + (1-\rho) \cdot ls(t) \quad (5)$$

with $\rho = 0.5$.

If *lsa* exceed a sieve, then *gs* is reset to the initial condition.

4. NORMALIZATION SCHEME

Based on the improvements proposed in the former section, the normalization algorithm employed in the experimental activity is characterized by performing:

- **Online MPEQ**, i.e. *global statistics* of time t-1 are used to normalize test utterance at time t. Voice presence probability is estimated by an MLP VAD.
- **Multimodal normalization:** *reference statistics* do not have a single component $rs = (\mu_{n,x}, \mu_{s,x}, \sigma_{n,x}, \sigma_{s,x})$ but a set of components $RS = \{rs_1, rs_2, ..., rs_k\}$ and the test utterance is normalized towards the nearest component. KLD is used with $\xi = 0.5$.
- **Distance controlled normalization**: normalization is applied only if the *global statistics* are sufficiently distant from each of the *reference statistics* components.
- **Switch context detection**: distance between *global statistics* and a short-term average of *local statistics* is used to detect a context switch (change of channel, noise, speaker), causing a reset of *global statistics*.

The algorithm implementation is given by the following steps:

1. **Initialization:** *RS* is initialized with the statistics of the main components within the training set, and *gs* with the element of *RS* with higher prior probability.
2. **Distance computation**:

   let $rs_m \,|\, m = \arg\min_i D(gs, rs_i)$ be the *RS* component nearest to *gs* according to distance D.
3. **Equalization:** if $D(gs, rs_m) > SN_D$, perform on-line MPEQ normalization of each frame *y* of test utterance by applying equation (3) and computing voice presence probability P(s, y) with MLP VAD.
4. **Statistics update**
   a. Compute *ls*(y)
   b. Update *gs* following equation (4);
   c. If $D(gs, lsa) > SC_D$, perform context switch, resetting *gs* value to the element of *RS* with highest prior probability. SC_D stands for the threshold that controls context switch. Distance employed is KLD.

5. EXPERIMENTAL RESULTS

Experimental activity has been devoted to test the proposed normalization scheme. The recognition system used is the commercial Loquendo ASR system 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 [6]. Normalization is not applied in training but only in test. Three models, characterized by multimodal training conditions, have been used in the experiments:

1. Loquendo American English 16kHz microphone (*micro*): the *reference statistics* set RS contains 5 different components, estimated on the five main components of the training set.
2. Loquendo American English 16kHz Automotive (*auto*): smaller model, optimized for embedded hw and automotive noise: the *reference statistics* set *RS* contains 4 different components.
3. Loquendo American English 8kHz Telephonic (*telephone*): 4 *reference statistics* components.

5.1 Test corpora

The following test corpora have been employed in the experiments:

WSJ0 5K:

Standard SI_ET_05 test set with 8 speakers and 40 sentences per speaker has been used. Two channels are evaluated: WV1, Senheiser microphone (matched condition) and WV2, other microphones (mismatched condition). Vocabulary: 5K words, with standard trigram LM from Lincoln labs. The ASR model used is micro.

Safesound

It is a subset of the corpus collected inside the EU project Safesound that studied the possibilities of improving safety for ground and flight operations by the application of enhanced audio functions in the cockpit of an airplane. The corpus is described in [7]. The recognition task is a small vocabulary (240 words) command and control task. Commands include page selections, display changes and parameter settings.

A rule-based grammar is used to model the command structure. Our subset includes the 8 Italian pilots, speaking English as non-native speakers. Each of them uttered about 200 sentences for a total of 1624 utterances. The ASR models used are *micro* and *auto*.

HIWIRE cockpit database

It is a noisy and non native English speech corpus for cockpit communications [8]. It includes short vocal sentences in English, corresponding to aeronautic commands. It includes 81 non-native speakers from 4 countries.

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 statistics for normalization purposes. The ASR model used is micro.

SpeechDatCar Italian (CH0) with artificial mismatch:

It is a subset of SpeechDatCar Italian (CH0) with artificial generated mismatch conditions. Four mismatch conditions have been generated with Sox to simulate typical problems:

1. **Attenuation**: volume has been reduced to 15% of the original one. It is still perceivable, but creates problems to ASR;
2. **Saturation**: the signal has been saturated. Still perfectly understandable; no problems for humans, but degraded performances for ASR.
3. **Filtering**: A microphone mismatch has been simulated by a band-pass filter that has attenuated low (< 500 Hz) and high (> 2200 Hz) frequencies;

The ASR model used is *telephone*.

5.2 Test Results

Results are collected in the following tables:

WSJ0 5K	NO PEQ	Online MPEQ	
microphone		$SN_D = 0.1$	$SN_D = 3.0$
Sennheiser	92.4	91.8	92.4
2nd microphone	73.9	76.4	75.9

Table 1. Word Accuracy results for WSJ0

Safesound	NO PEQ	Online MPEQ $SN_D = 3.0$	*Err. Red.*
Model: micro	84.0	91.3	*45.62%*
Model: auto	72.9	84.4	*42.40%*

Table 2. Word Accuracy results for Safesound subset

HIWIRE Cockpit non-native corpus			
Model: auto	NO PEQ	Online MPEQ $SN_D = 3.0$	*Err. Red.*
Clean	91.1	92.4	*14.6%*
Low Noise	79.0	81.4	*11.4%*
Mean Noise	67.6	70.6	*9.3%*
High Noise	28.0	30.0	*2.7%*
Average	66.4	68.7	*6.8%*

Table 3. Word Accuracy results for Hiwire cockpit non-native noisy corpus

SpeechDatCar Ch0 with artificial mismatch	NO PEQ	Online MPEQ Retrained model	Online MPEQ Not retrained model $SN_D = 3.0$
original	97.1	97.0	96.8
attenuated	84.1	90.4	88.6
saturated	78.9	87.6	85.6
filtered	84.5	89.7	88.9
Average	82.5	89.2	87.7
Average E.R.	-	*49.0*	*40.2*

Table 4. Word Accuracy results for SpeechDatCar Italian CH0 with artificial mismatch conditions

5.3 Discussion

WSJ0: Results show that Online MPEQ compensates channel mismatches: in fact it does not operate on the matching condition (a component of Sennheiser data from wsj0-1 is present in the training set) but operates on the other microphones data, with a 10% Error Reduction (E.R.). Two SN_D sieves are tested: with $SN_D = 0.1$ normalization is always performed, and the result on the 2nd microphone is better, but degradation on Sennheiser appears; with $SN_D = 3.0$ Sennheiser performance is maintained and improvement on 2nd microphone is only slightly inferior.

Safesound: It is the database for which the best results are obtained. In fact this corpus presents an important degree of microphone and channel mismatch with respect to the training set conditions. In particular some speakers appear much attenuated. Online MPEQ compensates well these problems obtaining a large E.R. of more than 42%.

HIWIRE cockpit database: This corpus is not well suited to be treated by Online MPEQ as the problems here are mainly non-native speakers and additive noise. Notwithstanding this, significant improvements are obtained especially in clean and low noise conditions.

SpeechDatCar Italian (CH0) with artificial mismatch: This corpus has been appositely designed to test Online MPEQ capabilities to deal with different kinds of channel mismatch. This test compares two Online MPEQ cases: the case where a retraining of ASR model has been done, using Online MPEQ also on the training set, and the more realistic case where the models cannot be retrained and Online MPEQ is applied only in test. Although the first case performances are better, as well known from previous literature [3][4], the second case still provides a large improvement (E.R. = 40.2%). In practice only the second kind of normalization can be deployed, as the models of a commercial ASR cannot be easily retrained.

6. CONCLUSIONS

An evolution of Memory PEQ normalization has been proposed. The evolution has the following desirable features:

a) It works online in real-time without introducing a delay in the recognition process;

b) It does not require a retraining of the ASR models;

c) It is not intrusive on utterances that already match well the conditions present in the training set;

d) It keeps explicitly into account the multiplicity of conditions present in the huge training sets that characterize modern ASR systems.

The presented results show good improvements in a variety of mismatched conditions, while performances in matched conditions are preserved.

REFERENCES

[1] Sirko Molau, Daniel Keysers, Hermann Ney. "Matching training and test data distributions for robust speech recognition", *Speech Communication 41* (2003) 579-601.

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

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

[4] L. Garcia, R. Gemello, F. Mana, J. Segura. "Progressive Memory-based Parametric Non-Linear Feature Equalization". *Proceedings of Interspeech 2009*, Brighton.

[5] Roberto Gemello, Franco Mana and Renato De Mori. "Non-linear estimation of voice activity to improve automatic recognition of noisy speech", *Proceedings of Interspeech 2005*, Lisbon

[6] D. Albesano, 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.

[7] Judith Kessens "Non-native Pronunciation Modeling in a Command & Control Recognition Task: A Comparison between Acoustic and Lexical Modeling" *ISCA Workshop on Multilingual Speech and Language Processing (MULTILING 2006)*, Stellenbosch, South Africa

[8] A. Potamianos, D. Fohr, I. Illina, P-A. Breton, V. Clot, R. Gemello, M. Matassoni, P. Maragos, J.C. Segura, T. Ehrette. "The hiwire database, a noisy and non-native english speech corpus for cockpit communications". *http://www.hiwire.org/*