On the Effect of the Language in CMS Channel Normalization

Dirceu Gonzaga da Silva, José A. Apolinário Jr., and Charles B. de Lima
[†]IME–DE/3, Rio de Janeiro–RJ, Brazil

Abstract— **This paper presents a modification in a technique of channel normalization widely known as Cepstral Mean Subtraction (CMS). This modification is based on the introduction of language dependent phonetic modification. A careful investigation using Brazilian Portuguese was carried out showing that it is possible to improve the CMS channel identification through a constant vector, associated with the language, obtained from an estimation of the mean cepstral coefficients from clean speech signal over time. As a consequence of better channel estimation, better features normalization is attained. Computer simulations were carried out with cepstral coefficients extracted from Mel-scale in a speaker identification experiment where the proposed technique, in some cases, improved the recognition rate on the top of the CMS good results.**

I. INTRODUCTION

ONE of the main problems reported in automatic speech and speaker recognition technology has been the different recording environments used in training and testing corpora. This mismatch may be caused by different transmission channels, different acoustic environment, and/or different microphones. Recent research has shown that, for the case of no mismatch and clean data, speech recognition systems have reached a good performance or, equivalently, they present very small error rates. Nevertheless, for the case where there is mismatch between training and testing signals, the performance drops significantly [1]. Nowadays, there is a growing interest and consequent increasing research effort in channel normalization techniques, which try to compensate for these *channel* distortions.

Channel normalization is usually carried out in some specific features extracted from speech signal. The mostly used features for this purpose are the ones based on the concept of homomorphic deconvolution [2] such as cepstral coefficients, which has the following formulation. Let $y(t)$ be the result of the convolution between signal $s(t)$ and $h(t)$, the channel impulse response, such that

$$y(t) = s(t) * h(t) \tag{1}$$

Once this signal is digitized and the Discrete Fourier Transform (DFT) is applied to each frame[1], the convolution in (1) results in a multiplication on the frequency domain.

$$|Y_{k,i}| = |S_{k,i}||H_k| \tag{2}$$

where k is the DFT index and i is the frame index. In order to simplify the notation, we will drop the DFT index and use $\mathbf{Y_i}$, $\mathbf{S_i}$, and $\mathbf{H}$ to denote the DFT vector (containing all DFT coefficients) for each frame.

Aiming the removal of the channel influence by subtracting its components, the logarithm function is applied to both sides of the previous equation such that a multiplication turns into an addition.

$$\log|\mathbf{Y_i}| = \log|\mathbf{S_i}| + \log|\mathbf{H}| \tag{3}$$

It is now possible to retrieve one information if we know the other one. An inverse transform is then used in (3) and the result, in the cepstral domain (DFT cepstrum in this case), is

$$\hat{\mathbf{y}}_\mathbf{i} = \hat{\mathbf{s}}_\mathbf{i} + \hat{\mathbf{h}} \tag{4}$$

where $\hat{\mathbf{y}}_\mathbf{i}$, $\hat{\mathbf{s}}_\mathbf{i}$, and $\hat{\mathbf{h}}$ are the cepstral coefficients vector for each frame of the distorted speech, the clean speech, and the channel, respectively.

Cepstral Mean Subtraction (CMS) [4], [5], also known as Cepstral Mean Normalization (CMN), is one of the most widely used schemes for channel normalization. This technique is based on the removal of the DC level obtained from the time evolution of the cepstral coefficients. This temporal mean is a rough estimate of the transmission

The authors are with the Department of Electrical Engineering, Instituto Militar de Engenharia, Praça General Tibúrcio, 80, Urca, 22.290-270, Rio de Janeiro–RJ, Brazil, Phone: +55 21 2546 7030 Fax: +55 21 25467039. E-mail: dirceu@epq.ime.eb.br, apolin@ieee.org, and cborges@epq.ime.eb.br. The authors thank FAPERJ and CAPES for partially funding of this work.

[1]The speech signal is divided in overlapping frames and each frame is windowed before applying the DFT. It is also assumed that the window size is around 4 times the channel impulse response [3].

channel or microphone response. Nevertheless, its far and wide use comprises both speech and speaker recognition [5], [6].

In [7], a modification in the CMS was proposed in order to compensate for the distortions introduced by the poles (of the all-pole filter in the LPC model) in the channel identification problem. This technique was known as *pole-filter* and presented a smoothing property on the peaks generated by those poles. In [6], two algorithms used to improve the CMS normalization were analyzed: one through *log-DFT mean normalization* which shows that the blind channel identification carried out by the CMS is suboptimal and a second using second order statistics with the help of a Hidden Markov Model (HMM) for the channel identification. The use of the second order statistics improves the normalization with the expense of a higher computational complexity.

In order to have the cepstral mean being admitted as a fair estimate of the channel, CMS assumes that clean speech cepstral mean tends to zero [8]. The goal of this paper is the proposition that this is not exactly true but depends on the language. We, therefore, propose a modification where the cepstral mean of the language is taken into account for a more reliable estimate of the channel. As a result, a more effective features normalization is accomplished.

This paper is divided as follows. In Section II, the formulation of the CMS and its basic problems are reviewed. Section III introduces the proposed method which tries to compensate for the effect of the language in the channel estimation. The simulation results of a speaker identification experiment are presented in Section IV followed by a few conclusions.

II. CMS Channel Normalization

Cepstral Mean Subtraction is a features normalization technique based on channel blind identification. Therefore, any previous knowledge about the signal features can bring some advantage to the algorithm. Let us first consider a frame mean in the cepstral domain given by

$$\bar{\hat{\mathbf{y}}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{y}}_i = \frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{s}}_i + \hat{\mathbf{h}}) \tag{5}$$

where N is the number of speech frames.

In many cases, a previous knowledge of a given speech signal is the assumption that the channel is time invariant as indicated above by the lack of subscript i in $\hat{\mathbf{h}}$. This means that, given the time evolution of a cepstral coefficient, the channel affects only its DC level. With this assumption, (5) can be written as

$$\bar{\hat{\mathbf{y}}} = \hat{\mathbf{h}} + \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{s}}_i \tag{6}$$

or,

$$\bar{\hat{\mathbf{y}}} = \hat{\mathbf{h}} + \bar{\hat{\mathbf{s}}} \tag{7}$$

According to [8], if the speech signal is balanced in terms of voiced, unvoiced, and plosive sounds, the cepstral mean tends to zero or $\bar{\hat{\mathbf{s}}} \rightarrow \mathbf{0}$ such that $\bar{\hat{\mathbf{y}}} \approx \hat{\mathbf{h}}$.

Once obtained the channel estimate, we can go further to normalization with the following subtraction

$$\hat{\mathbf{s}}_i \approx \hat{\mathbf{y}}_i - \bar{\hat{\mathbf{y}}} \tag{8}$$

Nevertheless, CMS presents two basic problems. The first one is the fact that the assumed balance hardly occurs. Moreover, this balance is expected to vary from language to language. The second problem concerns the subtraction of the cepstral mean itself which will not only remove the effect of the channel but anything constant and common for all speech frames. This means that we are loosing information. This problem was addressed in [9], where the variance of the signal was analyzed before and after CMS, showing that the speakers variance is reduced after CMS.

III. The Proposed Method

This section develops an approach which tries to improve channel normalization from the assumption that $\bar{\hat{\mathbf{s}}}$ does not tend to zero. We could name it a *Language Dependent Modified CMS*.

From the right side of (7), we note that if $\bar{\hat{\mathbf{s}}}$ is not zero, as expected by the CMS, the channel estimate will be biased. According to our experiments, this assumption ($\bar{\hat{\mathbf{s}}} \approx \mathbf{0}$ was not true for Portuguese. Computer experiments drove us to the conclusion that $\bar{\hat{\mathbf{s}}}$, the clean speech cepstral mean, remains constant for a particular language and a pre-defined sex (male or female). This assumption implies that, for each speaker, the cepstral mean tends to a constant value and also that these constants, for different speakers, present small variance.

In [6], it was shown that for the English language, the elements of $\bar{\hat{\mathbf{s}}}$ present small variance in channel identification if estimated over a minimum period of time. This fact supports our basic assumption for the English case. For the Portuguese case, two experiments were conducted in order check our claim of constant mean.

In the first experiment, the coefficients were extracted from a 2 minutes, reasonably clean (lab conditions),

speech signal recorded by a male speaker. The estimate was obtained via $\mathbf{m}_t = \frac{1}{t}\sum_{i=1}^{t} \hat{\mathbf{s}}_i$. Fig. 1 depicts the temporal evolution of this estimate for the first four Mel cepstral coefficients (MelC) [10].

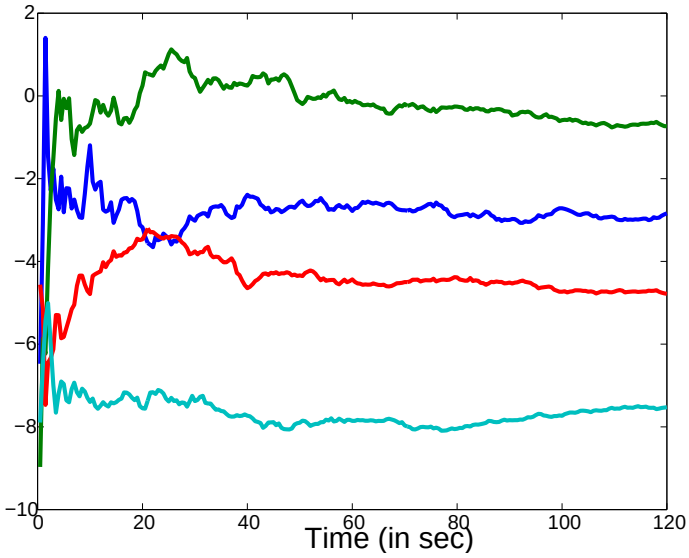

Fig. 1. Evolution of the estimates of the first four Mel coefficients.

Theoretically, the mean is obtained as $N \to \infty$. However, we can see in the figure that after some time – different for each coefficient – the estimation of the mean value of that coefficient tends to a constant. Other tests were carried out with several speakers and, in all cases, the results were similar.

The second experiment concerning the claim of a language dependent constant clean speech cepstral mean ($\mathbf{m}$) implies that this vector is representative for any speaker of the same sex and same language. Fig. 2 shows the results of two curves: the one with small squares corresponds to vector $\mathbf{m}$ estimated from around 30 minutes of undistorted Portuguese, with different phrases from [11], spoken by 12 male speakers. The second curve comes from 15 other male speakers, each one speaking around 30 seconds of undistorted Portuguese. This curve contains the mean plus minus standard deviation shown with vertical bars marked with "X" at the extremities. From Fig. 2, we can see that the mean estimated from 15 speakers is very close to the one obtained with 30 minutes. This suggests that $\mathbf{m}$ corresponds to a representative constant for speech signals for a given language (and sex) and can be used as a reasonable approximation of $\bar{\hat{\mathbf{s}}}$ for shorter signals. The experiments have shown that this mean vector is speaker independent and that they can be understood as a linguistic feature of a language.

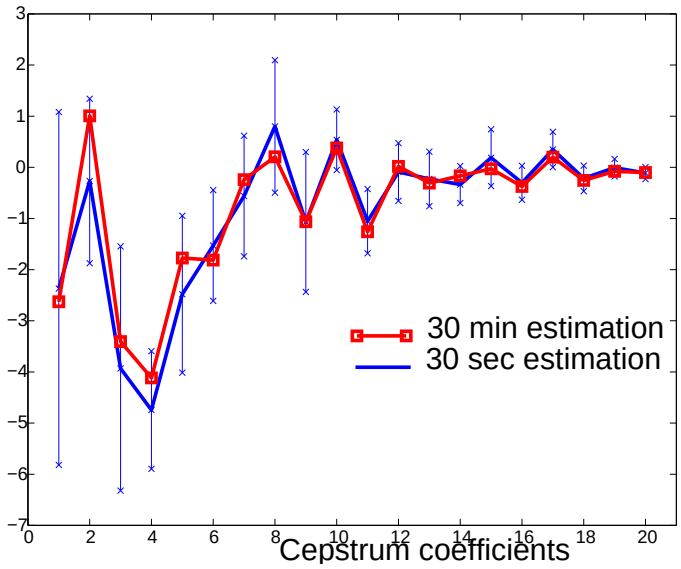

Fig. 2. Comparing Mel cepstral coefficients from 2 distinct estimations.

Assuming the term $\bar{\hat{\mathbf{s}}}$ being approximately equal to the cepstral coefficient mean estimated from a clean signal, $\mathbf{m}$, we can obtain a more reliable channel estimate, according to $\bar{\hat{\mathbf{y}}} - \mathbf{m} = \hat{\mathbf{h}} + \bar{\hat{\mathbf{s}}} - \mathbf{m}$ such that if $\bar{\hat{\mathbf{s}}} \approx \mathbf{m}$ then

$$\hat{\mathbf{h}} \approx \bar{\hat{\mathbf{y}}} - \mathbf{m} \tag{9}$$

In the previous section, it was mentioned that for blind channel estimation any prior knowledge can improve the result. We now show that this prior knowledge of the language characteristic given by $\mathbf{m}$, leads to an improved estimation of the channel. In order to test channel estimation, two telephone channels were used. One follows ITU Recommendation G.151 and will be designated Channel A. The other one is a digital model of a continental poor voice channel, designated Channel B. Both channels can be observed in Fig. 3 as well as the results of a blind channel estimation using conventional CMS, the proposed modified CMS, and what we have named Best Possible Estimation (viz BPE, the non-real case where we have both clean and distorted signals and obtain the channel estimation by subtracting the cepstrum of the clean speech from the cepstrum of the distorted signal according to (4)). In this particular example, 20 LPC cepstral coefficients (LPCC) [12], were used. The test consisted of estimating the frequency response of these two channels with 2 minutes of speech convolved with A and B. The (Brazilian Portuguese/male) mean vector used was obtained with the 30 minutes estimation previously described.

From Fig. 3, we can observe that the estimation through the proposed modified CMS is superior than that from the

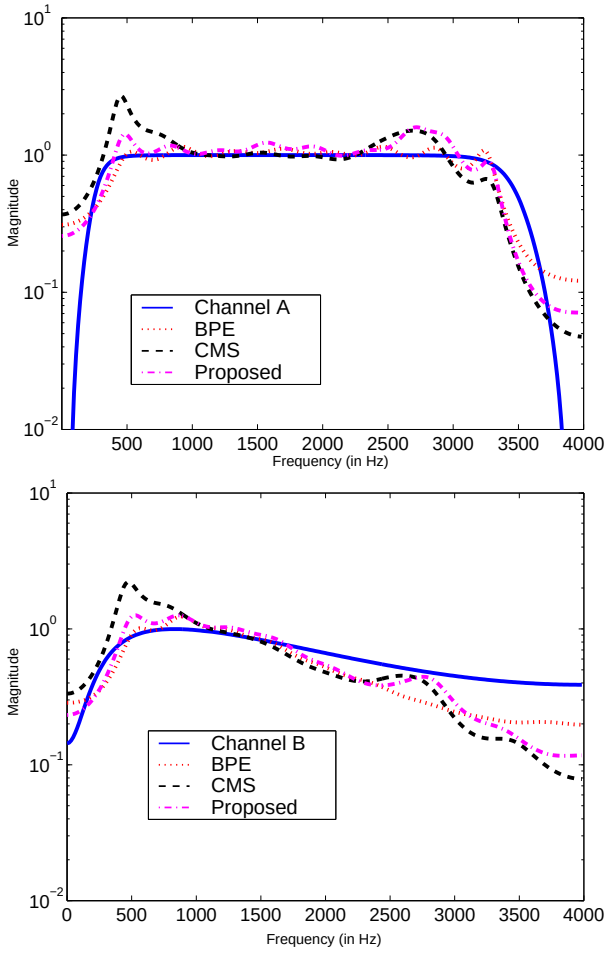

Fig. 3. Channel estimation with conventional CMS, proposed modification, and the theoretical best possible estimation.

The output result for each possible speaker (Spk k), $k \in \{1$ to $K\}$ is the total distance D_k given by:

$$D_k = \frac{1}{N}\sum_{i=1}^{N} \min_{1\le j\le M} d(a_i, b_j) \tag{10}$$

where k corresponds to the speaker index, M is the number of centroids, and N is the number of windows of the test signal.

For the computation of $d(a_i, b_j)$, the Euclidean distance was used. This distance (as well as the VQ itself) was chosen for its simplicity and the fact that our goal was the comparative analysis of the efficiency of the proposed technique, not obtaining the best possible recognition rate.

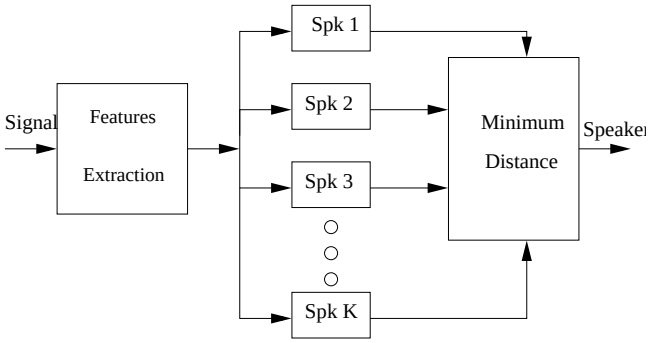

Fig. 4. Identification System using VQ.

conventional CMS procedure.

IV. Experiment in Speaker Identification

Automatic Speaker Recognition (ASR) is a generic term concerning the task of discriminating people based on their speech features. ASR can be classified according to their task as *Speaker Identification* and *Speaker Verification.* In this section, we will address the text independent speaker identification problem which corresponds to the classification of an utterance as belonging to one specific speaker from a—closed in our case—set of reference speakers.

A. Decision System Used

Vector Quantization (VQ) applied to ASR was introduced in [13] and was the decision system chosen for the evaluation of the proposed compensation schemes. Each speaker codebook was obtained from 32 speech features groups extracted from each speaker utterance. For the training, it was used the LBG algorithm as described in [14]. The system identification scheme is shown in Fig. 4.

B. Corpus Used

In order to carry out simulations of the speaker identification using the proposed normalizations, a speech data base was recorded in the Speech Lab of the Instituto Militar de Engenharia (IME). This simple corpus was formed by recording speech from 50 male speakers, each one speaking 20 groups of 10 phrases proposed in [11].

The silence periods of the speech signals from these phrases were extracted and divided as follows:

- speech used for training and testing – we used the first 18 groups such that:
  - training speakers: 40 speaking 1min each;
  - test utterances: 474 recorded by the 40 speakers, each utterance with 30sec of speech.
- speech used for obtaining vector $\mathbf{m}$: the last 10 speakers were used and the last two groups of text read such that training and test utterances were different.

The training utterances were filtered through channel A and the test utterances through channel B.

C. Speaker Identification Results

Table I presents the error rate[2] obtained under the following conditions: without channel compensation, when CMS was used, and when the proposed method, the language dependent modified CMS, was used. Note, in the first column, the huge error caused by channel mismatch when VQ is used. Also note in this table that the configuration presented was concerning the frame rate: the frame size in milliseconds and the overlapping of adjacent frames. It is worth-mentioning that the feature used was the Mel Cepstral Coefficients (MCC) which in prior experiments carried out with the same corpus resulted in the lowest error rate when Mel, LPC, DFT, and PLP [15] were compared with no compensation scheme. In the speaker identification experiment described here, considering that we are using telephone channels, we have used only the filters (from the Mel scale filter bank) which central frequencies were inside the typical telephone bandwidth of $300 - 3400$Hz.

TABLE I
ERROR RATE IN % OF THE SPEAKER IDENTIFICATION USING VQ WITH NO COMPENSATION, CMS, AND THE MODIFIED CMS

Config.	No Compens.	CMS	Modif. CMS
20ms 50%	79.54	1.05	0.42
20ms 75%	79.96	0.42	0.42
40ms 75%	79.11	0.84	0.84
40ms 50%	81.01	1.05	0.84

V. CONCLUSIONS

This paper addresses CMS blind channel identification using phonetic information of the Brazilian Portuguese. Nevertheless, the technique proposed here is expected to be valid for any other language. It is shown that the language contains constant information in the mean cepstral coefficients obtained form speakers of the same sex. This information can be used to improve the channel estimation used in the CMS approach. It is worth mentioning that the proposed modification and the conventional CMS are comparable in terms of computational complexity, for they have the same number of multiplications.

It is important to emphasize that the results obtained so far are preliminary due to the following aspects:

- the corpus used is very small and we do not have an adequate Portuguese corpus available for research in speaker recognition. We are currently working in the development of a 1000 speakers data base;
- the language mean must be obtained from speech signals recorded in an acoustically isolated chamber and with high quality microphones;
- the main idea proposed here should be investigated in other languages.

[2] Rate between the number of wrong identifications and the total number of tests.

REFERENCES

[1] D. A. Reynolds. *Experimental Evaluation of Features for Robust Speaker Identification.* IEEE Trans. on Speech and Audio Processing, vol. 2, Nr. 4, Oct. 1994.

[2] A. V. Oppenheim, R. W. Schafer. *Digital Signal Processing.* Englewood Cliffs, NJ, Prentice-Hall, 1989.

[3] C. Avendano and H. Hermansky. *On the Effects of Short-Term Spectrum Smoothing in Channel Normalization.* IEEE Transaction on Speech and Audio Processing, Jul. 1997.

[4] B. Atal. *Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker identification.* JASA, 55, pp. 1304-1312, Jun. 1998.

[5] S. Furui. *Cepstral Analysis Technique for Automatic Speaker Verification.* IEEE Trans. on Acoust. Speech, and Signal Processing, 29, pp. 254-272, Apr. 1981.

[6] L. G. Neumeyer, V. V. Digalakis, and M. Weintraub. *Training Issues and Channel Equalization Techniques for the Construction of Telephone Acoustic Models Using a High-Quality Speech Corpus.* IEEE Transaction on Speech and Audio Processing, 2(4), pp. 590-597, Oct. 1994

[7] D. Naik. *Pole-Filtered Cepstral Mean Subtraction.* Proceedings of the ICASSP, pp. 157-160, 1995.

[8] R. J. Mammone, X. Zhang and R. P. Ramachandran. *Robust Speaker Recognition – A Feature-based Approach.* IEEE Signal Processing Magazine, pp.58-71, Sep. 1996.

[9] S. Kajarekar, M. Malayath, and H. Hermansky. *Analysis of Speaker and Channel Variability in Speech.* Proceeding of the Workshop on Automatic Speech Recognition, and Understanding, Keystone, CO, Dec. 1999.

[10] P. Mermelstein and S. B. Davis. *Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences.* IEEE Trans. on Acoust., Speech, and Signal Processing, 28(4), pp. 357-366, Aug. 1980.

[11] A. Alcaim, J. A. Solewicz, and J. A. Moraes. *Freqüência de Ocorrência dos Fonemas e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro.* Revista da SBrT, 7(1), pp. 23-41, Dec. 1992.

[12] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete Time Processing of Speech Signals.* Macmillan Publishing Company, New York, 1993.

[13] F. Soong, A. Rosemberg, B. Juang, and L. Rabiner. *A Vector Quantization Approach to Speaker Recognition.* AT&T Technical Journal, March 1987.

[14] Y. Linde, B. Buzo, and R. Gray. *An Algorithm for Vector Quantizer Design.* IEEE Transactions on Communications, Jan. 1980.

[15] H. Hermansky. *Perceptual Linear Predictive (PLP) Analysis of Speech.* J. of the Acoust. Soc. of Am., 87(4), pp. 1738-1752, Apr. 1990.