



On enhancing feature sequence filtering with filter-bank energy transformation in speaker verification with telephone speech

Claudio Garretón and Néstor Becerra Yoma

Speech Processing and Transmission Laboratory, Department of Electrical Engineering
Universidad de Chile, Santiago, Chile.
{cgarreto, nbecerra}@ing.uchile.cl

Abstract

In this paper a novel feature enhancing method for channel robustness with short utterances is employed. The transform reduces the time-varying component of the channel distortion by applying a band-pass filter along the filter-bank domain on a frame-by-frame basis. This procedure enhances the channel cancelling effect given by techniques based on feature trajectory filtering. The transformation parameters are defined employing relative importance analysis based on a discriminant function. In text-dependent speaker verification with telephone speech the transform leads to a reduction in the EER of 10.8%, and further improvements of 23.5% and 40% when combined with RASTA or CMN, respectively.

Index Terms: robust features, speaker recognition, text-dependent speaker verification, telephone speech.

1. Introduction

Robustness to the mismatch between training and testing channel conditions is one of the most important challenges faced by speech and speaker recognition systems in practical situations. In a telephone-based real application, due to operating and usability restrictions, the amount of available data to remove or reduce convolutional noise is limited. For instance, enrolling and verification procedures in text-dependent speaker verification (TD-SV) systems over the telephone network should be fast and efficient. As a consequence, a limited enrolling/testing data scenario leads to severe degradations in the accuracy of TD-SV engines.

The motivation of channel canceling techniques is to reach system performance observed in channel matched conditions by minimizing the requirements of extra data. The approaches that address the problem of channel mismatch can be clustered into two main areas [1]: feature compensation [2-4]; and, model adaptation [5-6]. The most widely accepted model for channel distortion corresponds to a cepstral or log-spectral bias that results from the following hypotheses: H1, the channel response is signal independent; and H2, the channel can be modeled as a linear filter. Based on hypotheses H1 and H2, current feature compensation methods estimate the original undistorted signal by removing a bias or low-frequency component in the cepstral or log-spectral domain [2-4]. Usually these approaches can dramatically reduce the error rate in channel mismatch condition but also show a significant efficacy lost when limited data is available. For instance, CMN attempts to remove the bias component in the cepstral domain, but its effectiveness is reduced with short utterances [7]. Moreover, bias removal methods based on the EM algorithm can also provide significant reductions in error rate [3]. Nevertheless, the EM algorithm is also sensitive to utterance length, usually requires a high computational load and does not remove convolutional distortion with complete

effectiveness. Those results must be due to the fact that the bias or low-frequency component in the feature domain can only account for a part of the channel effect. As a result, hypotheses H1 and H2 may lose validity. In fact, there are empirical evidences that strongly suggest that channel distortion depends on the input signal [8]. As a result, there is a time-varying component of the convolutional noise due to the channel distortion dependence on the speech signal. This component restricts the validity of hypotheses H1 and H2.

The estimation of the time-varying component of the convolutional distortion is a challenging task that has hardly been addressed in the literature. This distortion component is frame dependent and cannot be estimated with the information provided by a single frame. Some utterance-based techniques have been proposed to address the problem of channel distortion without making use of hypotheses H1 and H2. However, those approaches may lose their effectiveness when limited data is available. For instance, CMVN [9] has demonstrated to be an efficient alternative for restoring the clean speech features in some text-independent speaker verification (TI-SI) tasks, where verification utterances can be as long as hundreds of seconds. In contrast, TD-SV task employs short test utterances (shorter than 5 seconds), this situation leads to a dramatic reduction in the amount of available data and phonetic variability. Also, phonetic mismatch may occur between training and testing conditions. As can be seen in the results presented here, when CMVN is applied to a short utterance task the cepstral variance is unreliably estimated, which in turn leads to an inaccurate distortion cancelling and to an increase of the error rate.

Frequency filtering techniques have already been proposed to enhance the discrimination ability of HMM-based ASR systems [10-11]. As shown in [10], filtering along the log mel filter-bank energy (LFBE) vector can produce two main effects: a decorrelation of the sequence of features; and, a weighting of the cepstral coefficients. Despite the fact that those techniques may increase the recognition accuracy when compared with classical parameterizations (e.g. MFCC), they have not been employed to address the robustness to telephone channel mismatch conditions.

The transform employed in this paper [12] attempts to reduce the effect of the time-varying telephone channel distortion by applying a band-pass filtering along the LFBE feature vector on a frame-by-frame basis without considering hypotheses H1 and H2. This procedure is equivalent to de-emphasizing some components and enhancing others in the domain of the spectrum of LFBE features. As a result, the channel mismatch between models and test signals can be significantly reduced. Observe that the feature enhancement scheme used here applies a filter along the LFBE feature vector and complements the effect of standard "time-domain" methods that filter the trajectory of each feature.

In contrast to existing frequency filtering approaches, the scheme employed in this paper considers the use of a relative importance analysis (RIA) that makes use of an evaluation database recorded on several channel conditions to obtain the transformation parameters. In this paper a discriminative based objective function is used to apply RIA.

2. LFBE spectrum domain transformation

The scheme employed in this paper transforms the LFBE feature space in order to increase the system discrimination ability by reducing the time-varying component of channel distortion, independently of hypotheses H1 and H2. The main idea is to reduce those components in the spectrum of the LFBE vector that are more sensitive to channel distortion and preserve those components that contain mainly speech specific information. Consider that $Y_t[m]$ denotes the m th log Mel filter-bank energy at frame t . The discrete spectrum Z_t of Y_t , can be evaluated by employing a K -dimensional discrete Fourier transform, $DFT\{\}$:

$$Z_t = DFT\{Y_t\} \quad (1)$$

The filtering scheme can be applied in the spectrum of LFBE, Z_t , as a weighting function $G[k]$, $0 \leq k < K$. Consequently, the k th component of the filtered LFBE spectrum, $\hat{Z}_t[k]$, is:

$$\hat{Z}_t[k] = G[k] \cdot Z_t[k] \quad (2)$$

The filtered LFBE vector at frame t , \hat{Y}_t , is obtained by applying the inverse DFT transform, to (2). As a result, \hat{Y}_t is equal to the circular convolution of Y_t with the impulse response of G , g :

$$\hat{Y}_t = g \otimes Y_t \quad (3)$$

Notice that (3) allows applying the proposed filter along the feature frame in the LFBE domain. In this paper the TD-SV task employs cepstral features. Consequently, the convolution indicated in (3) between g and Y_t is applied as a previous stage before computing the cepstral parameters by using the discrete cosine transform.

3. Relative importance analysis for the definition of function G

This paper makes use of RIA [13] to define function G in (2) as proposed in [12]. RIA estimates the contribution of spectral components on the system performance. By doing so, it is possible to identify those components that are more sensitive to channel distortion and those ones that contain mainly speech specific information. If the former components are attenuated and the later ones are enhanced, the effect of the time-varying channel distortion will be diminished. RIA is employed to evaluate the robustness to convolutional noise of the spectral component k in the Z_t domain. Originally, RIA was proposed to make a straightforward use of a system performance measure (e.g. EER in SV and recognition accuracy in ASR) [13]. In contrast, in this paper the relative importance measure is computed using a discriminant function as proposed in [12]. An evaluation database (see Section IV) was employed to estimate the relative importance of each component in the LFBE spectrum domain. This database is composed of 40 speakers and each one was recorded with three different types of telephone handsets. As mentioned above, the TD-SV system used here makes use of cepstral features. Consequently, the discriminative function J

employed by RIA should also be estimated in the cepstral domain. A background Gaussian mixture model (GMM) is trained by employing an evaluation database composed of 40 speakers and a speaker-dependent GMM is generated for each speaker in the evaluation set by employing MAP adaptation [14]. By doing so, the correspondence of the Gaussians within each speaker-dependent GMM with those in the background GMM is preserved [14]. The discriminative function J used in RIA is defined here as:

$$J = \frac{1}{P} \sum_{p=1}^P \frac{1}{N} \sum_{n=1}^N \frac{\text{Inter-speaker variability at feature } n}{\text{Intra-speaker variability at feature } n} \quad (4)$$

$$= \frac{1}{P} \sum_{p=1}^P \frac{1}{N} \sum_{n=1}^N \frac{\sum_{s=1}^S (\mu_{s,p}[n] - U_p[n])^2}{\sum_{s=1}^S \sigma_{s,p}^2[n]}$$

where N denotes the number of cepstral features, $\mu_{s,p}[n]$ and $\sigma_{s,p}^2[n]$ are the mean and variance of feature n in Gaussian p of speaker s . Notice that diagonal covariance matrixes are used. S is the number of speakers and P is the number of Gaussians. In this paper P is made equal to 128. Also, $U_p[n]$ is the mean value of feature n associated to Gaussian p in the evaluation database. It is worth emphasizing that the discriminative function J defined in (4) is an intraclass-interclass dispersion ratio and can be interpreted as a measure of the class-mean separability on a least-square sense. Consequently, if a given transform is optimized by maximizing J , the discrimination ability should be increased.

Consider that function G in (2) is defined by low and high cut-off frequencies, k_L and k_H , respectively. If K denotes the dimension of the DFT analysis in (1), the domain of G can be defined by the first $K/2 + 1$ components ($K=16$ in this paper). Then, $0 \leq k_L \leq K/2$ and $0 \leq k_H \leq K/2$. $G[k]$ is defined as:

$$G[k] = \begin{cases} w_L & k < k_L \\ 1 & k_L \leq k \leq k_H \\ w_H & k_H < k \end{cases} \quad (5)$$

where w_L and w_H are the gains in the low and high stop bands, respectively. The relative importance measure associated to component k , $R(k)$, $0 \leq k \leq K/2$, is:

$$R(k) = \frac{1}{K/2} \left\{ \begin{aligned} & \sum_{k_1=0}^{k-1} [J(k_1, k) - J(k_1, k-1)] \\ & + \sum_{k_2=k+1}^{K/2} [J(k, k_2) - J(k+1, k_2)] \end{aligned} \right\} \quad (6)$$

where $J(k_1, k_2)$ is the proposed discriminative function defined in (4), estimated with the evaluation data when the proposed filtering scheme is applied according to (3) with parameters $k_L=k_1$, $k_H=k_2$ and $w_L=w_H=0.0$ in (5). According to [13], $R(k)$ is obtained by differentiating the surface of $J(k_1, k_2)$ with respect to a given component k , where k can be either a lower or an upper bound. Then, $R(k)$ in (6) corresponds to the average derivative by considering all the upper or lower bounds depending if k is a lower or upper bound in (5), respectively. This analysis can be interpreted as the average improvement resulting from the inclusion of a certain component k and explains the summations in (6). As shown in Fig. 1, discriminative function $J(k_1, k_2)$ can give a higher value than the baseline system (i.e. $G[k]$ is not applied, or $k_1=0$ and $k_2=K/2$) when some components in the LFBE spectrum domain are removed. The estimated $R(k)$ with $0 \leq k \leq K/2$ is presented Fig. 2. Three configurations were used: the baseline system without trajectory filtering of features; the baseline system with CMN; and, the baseline system with

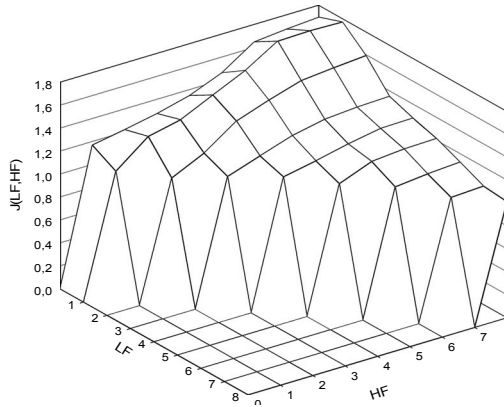


Figure 1: Discriminative function $J(k_1, k_2)$ vs. (k_1, k_2) in $G[k]$.

RASTA. As can be seen in Fig. 2, some components in the LFBE spectrum domain provide much higher relative importance measure $R(k)$ than others. In fact, some components show a negative $R(k)$, which in turn suggests that some LFBE spectral bands should be attenuated or removed to increase the discrimination ability. Actually, the relative importance analysis in combination with the discriminative function J defined in (4) suggests that the lower and upper bounds of $G[k]$, k_L and k_H , respectively, can be defined when $R(k) < 0$ or $R(k) \equiv 0$. As a consequence of this analysis, filter G in (5) is employed with the following cut-off frequencies: $k_L=1$ and $k_H=6$ without trajectory filtering of features and with CMN; and, $k_L=0$ and $k_H=6$ with RASTA. After k_L and k_H have been defined the corresponding gains w_L and w_H defined in (5), respectively, can be estimated by tuning. Therefore, the tuning procedure adopted in this paper can be considered guided by a previous data-driven analysis. Observe that there is not a straightforward numerical relation between the RIA results in Fig. 2 and the parameters of filter $G[k]$ in (5). Concluding, this result validates the use of the LFBE spectral domain to obtain a concise representation of the channel distortion effect. Actually, Fig. 2 clearly shows that the distortion effect is much more acute in some LFBE spectral bands than others.

4. Experiments

The feature enhancing scheme employed here is tested with the TD-SV system of the Speech Processing Laboratory, University of Chile. The speech signals were divided in 25 ms frames with 12.5 ms overlapping; each frame was processed with a Hamming window. The band from 300 to 3400 Hz was covered with 14 Mel DFT filters and at the output of each channel the logarithm of the energy was computed. Next, the feature filtering approach employed in this paper [12] was applied as in (3) before the cepstral transform. Finally, the frame energy plus ten static cepstral coefficients, and their first and second time derivatives were estimated. The HMMs were trained with the Viterbi algorithm. Each triphone was modeled with a three-state left-to-right HMM topology without skip-state transition, with one multivariate Gaussian density per state in speaker-dependent and background speaker models. Diagonal covariance matrices were employed. In the verification procedure each utterance is processed using the forced-Viterbi algorithm in order to estimate the log-likelihood. The output score of the system is a log-likelihood ratio, and the normalizing term is the averaged log-likelihood of the models of a cohort of the background speaker set.

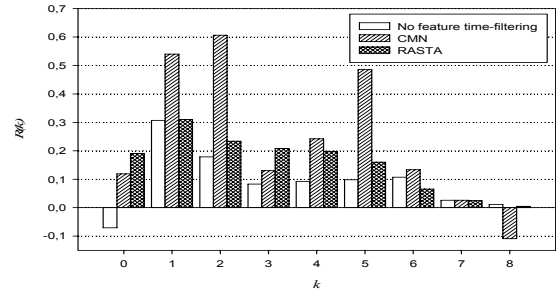


Figure 2: $R(k)$ vs. component k of the LFBE spectrum domain with and without filtering of time trajectories of features.

Results presented here were achieved with a telephone version of YOHO database [15]. In this paper a subset of 70 speakers (40 males and 30 females) was employed. The speakers were divided as follows: 40 background impostor speakers (20 males and 20 females) to train the background models; and, 30 testing speakers (20 males and 10 females) were used in verification attempts. For each selected speaker, one 24-utterance enrollment session was considered. Four verification sessions per testing speaker were employed, where four utterances were selected per session. Each utterance was recorded on a real landline telephone call by employing a PC speaker and telephone microphone acoustic coupling. Seven telephone handsets were used (hset1, hset2, ..., hset7). The signals were sampled at 8 kHz and 16 bits per sample. The telephone database was post-processed with a FFT equalizer filter in order to compensate for the spectral distortion caused by the speaker and the soundcard. Handset hset1 was labeled as the reference or “matched” channel. The telephone database was divided in three subsets: Y1 (evaluation database), composed of the enrolling utterances from the 40 background speakers recorded with handsets hset2, hset3 and hset4; Y2, composed of the enrolling utterances from the 40 background speakers recorded with hset1; and Y3, composed of the enrolling and testing utterances from the 30 testing speakers. As explained in Section III, Y1 was used to estimate $J(k_1, k_2)$ and $R(k)$. Database Y2 was used to compute the impostor hypothesis score on verification attempts as described in [7]. The clients’ models were generated with Y3 by making use of the enrolling utterances recorded with hset1. The verification attempts were performed by employing testing utterances recorded with hset5, hset6 and hset7. Consequently, false rejection was estimated with 3 handsets x 30 speakers/per handset x 16 verification signals per client = 1440 utterances. False acceptance curves were obtained with 3 handsets x 29 impostors/per handset x 6 verification signals/per impostor x 30 speakers = 15660 experiments. Observe that, as suggested by some authors, the ratio between client and impostor attempts is approximately equal to 1 to 10. The baseline system gives an EER equal to 2.71% with matched channel conditions (hset1). When the whole testing database Y3 is used (hset5, hset6 and hset7) the baseline EER is equal to 4.17%. Observe that the baseline system does not use any trajectory filtering of features.

5. Discussion

As mentioned in Section III, Fig. 1 shows that the suppression or attenuation of some components in the LFBE spectrum domain can lead to higher $J(k_1, k_2)$ than the baseline system in a distinguishable region (k_1, k_2) . Figure 2 shows the relative importance measure with and without trajectory filtering of features. As can be seen in Fig. 2, some components of the LFBE spectrum provide more discriminant information than

Table 1: EER(%) for the baseline system, feature transform, CMN, RASTA, CMN combined with feature transform, RASTA combined with feature transform and CMVN.

	Baseline	RASTA	CMN	CMVN
EER(%)	4.17	3.28	2.61	5.72
EER(%) with feature transform	3.72	3.19	2.50	—

others in the presence of channel mismatch. The lowest values of $R(k)$ correspond to the lowest and highest spectral components, $k=0$, and $k=7$ or $k=8$, respectively, without trajectory filtering of features. However, when CMN and RASTA are used the relative importance function increases in the lowest spectral component ($k=0$). This can be due to the fact that the classical methods based on trajectory filtering of features also reduce the channel distortion in the lowest component of the LFBE spectrum domain. As a result, a lower attenuation in the low stop band than the one without trajectory filtering of features should be required. Consequently, $G[k]$ defined in (5) was evaluated with the following configurations: $k_L=1$ and $k_H=6$, $w_L=0.4$ and $w_H=0.0$ without trajectory filtering of features; $k_L=1$ and $k_H=6$, $w_L=0.8$ and $w_H=0.0$ with CMN; and, $k_L=0$ and $k_H=6$, and $w_H=0.0$ with RASTA. According to Table 1 and Fig. 3 the LFBE feature filtering can lead to reductions in EER as high as 10.8% when compared with the baseline system. Also, when combined with RASTA and CMN the employed scheme can lead to further reductions of 2.7% and 4.2% in EER, respectively. In addition, when the filtering scheme is applied in combination with RASTA and CMN, overall reductions in EER equal to 23.5% and 40% are achieved when compared with the baseline system without the trajectory filtering of features, respectively. Also, Table 1 shows that the use of CMVN increases the error rate in 37.2% when compared with the baseline system. This must be due to the inaccurate estimation of the cepstral variance when limited data is available in training and testing procedures.

6. Conclusions

In this paper a novel frame-by-frame based channel robust feature transform was employed in a text-dependent speaker verification task with telephone speech. The transformation is applied as a band-pass filter along the log Mel filter-bank energy domain to reduce the effect of the time-varying component of channel distortion, which in turn is generated by the dependence of the channel response on the input speech signal. The transform is applied on a frame-by-frame basis. Consequently, a frame-dependent compensation is achieved. Also, in contrast to conventional feature filtering techniques, the current paper does not consider the channel distortion as an additive constant or slow-varying component in the trajectory of log-energy or cepstral parameters. The filter is defined with a relative importance analysis in combination with a discriminative function based on intra-speaker/inter-speaker dispersion ratio. The results presented here show that the spectrum of the log Mel filter-bank energy is a domain that provides a concise representation of the channel distortion effect. Experiments with a TD-SV system show that the feature filtering method can achieve significant improvements in EER when applied alone and combined with conventional methods based on trajectory filtering of features such as CMN and RASTA. Finally, the combination of the filter-bank energy transformation with other channel removal approaches can be proposed as future research.

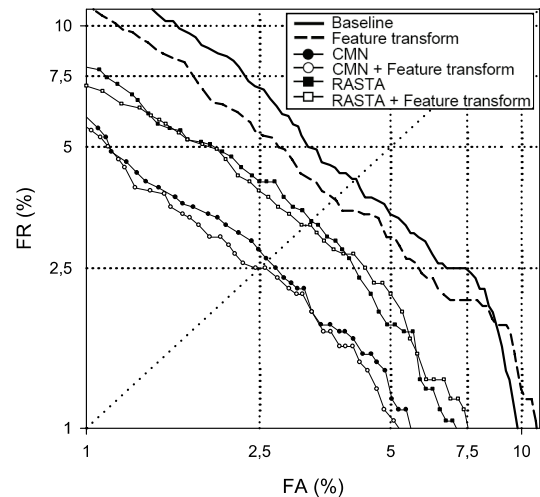


Figure 3: DET curves for the baseline system, feature transform, CMN, RASTA, CMN combined with feature transform, and RASTA combined with feature transform.

7. Acknowledgements

This work was funded by Conicyt-Chile under grants Fondecy D051-10243 and Fondecy 1070382 / 1100195.

8. References

- [1] M.W. Mak et al., "Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification," *EURASIP J. on Applied Signal Proc.*, 4, pp. 452-465, 2004.
- [2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on ASSP*, 29(2), pp. 254-272, 1981.
- [3] M.G. Rahim and B.H. Juang, "Signal bias removal by maximum likelihood estimation for robust telephone speech recognition," *IEEE Trans. on SAP*, 4(1), pp. 19-30, 1996.
- [4] Z. Tufekci, "Convolutional bias removal based on normalizing the filterbank spectral magnitude," *IEEE Signal Processing Letters*, 14(7), pp. 485-488, 2007.
- [5] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comp. Speech and Lang.*, 9(4), pp. 806-814, 1995.
- [6] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. on SAP*, 2(2), pp. 291-298, 1994.
- [7] L. Wang et al., "Robust distant speech recognition by combining position-dependent CMN with Conventional CMN," *Proc. ICASSP 2007*, pp. 817-820, 2007.
- [8] D.A. Reynolds et al., "The effects of telephone transmission degradations on speaker recognition performance," *Proc. ICASSP'95*, pp. 329-332, 1995.
- [9] Rong Zheng et al., "A comparative study of feature and score normalization for speaker verification," *Lecture Notes in Computer Science 3832*, pp. 531-538, 2005.
- [10] C. Nadeu et al., "Time and frequency filtering of filter-bank energies for robust HMM speech recognition," *Speech Communication*, 34 (1-2), pp. 93-114, 2001.
- [11] H.Y. Jung, "Filtering of filter-bank energies for robust speech recognition," *ETRI Journal*, 26 (3), pp.273-276, 2004.
- [12] C. Garreton, N.B. Yoma and M. Torres, "Channel robust feature transformation based on filter-bank energy filtering," *IEEE Trans. on ASLP*, 18(5), pp. 1082-1086, 2010.
- [13] N. Kanedera et al., "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, 28(1), pp. 43-55, 1999.
- [14] D.A. Reynolds et al., "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, 10(1), pp. 19-41, 2000.
- [15] Campbell J. and Higgins A., "YOHO speaker verification," *LDC*, Philadelphia, 1994.