Telephone Channel Compensation in Speaker Verification Using a Polynomial Approximation in the Log-Filter-Bank Energy Domain

Claudio Garreton and Nestor Becerra Yoma

***Abstract*—This correspondence presents a novel feature-space channel compensation technique that models the convolutional distortion in the log-energy mel-filter domain by means of a polynomial approximation. The proposed parametric distortion model generates appropriate constraints in the spectral domain that help to improve the channel cancelling estimation with limited data. In a text-dependent speaker verification task, the polynomial-based channel estimation scheme can lead to reductions in equal error rate (EER) as great as 22% and 8% when compared with the baseline system and with the standard cepstral bias removal approach, respectively, with no significant increase in computational load.**

***Index Terms*—Channel robustness, speaker recognition, text-dependent speaker verification.**

I. Introduction

In commercial applications on the telephone, two main restrictions are imposed on speaker verification (SV) technology: the amount of data for training and testing procedures is limited due to operating and usability restrictions; and speech signals are distorted by the communication channel that is composed of the telephone line and microphone handset. Limited data leads to poorly trained models or to inaccurate estimation of convolutional noise, and a consequent inaccurate compensation in testing, which in turn leaves SV systems extremely vulnerable to channel mismatch conditions.

Previous attempts to resolve the channel mismatch problem have had the goal of reaching the error rate that is observed in matched conditions while minimizing the requirements of estimation data. They can be clustered in two main areas [1]: feature compensation [1]–[5]; and model adaptation [6]–[8]. In both cases 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 time-invariant filter. Based on those hypotheses, state of the art feature compensation or model adaptation methods consider channel distortion as a bias in the Mel frequency cepstral coefficient (MFCC) or log filter-bank energy (LFBE) domains. If $x(t)$, $n(t)$ and $h(t)$ represent, respectively, the clean speech signal, additive noise, and the impulse response of the linear time-invariant filter that models the channel distortion [5], [9], the observed distorted signal $y(t)$ is modeled as

$$y(t) = [x(t) + n(t)]^{*} h(t). \tag{1}$$

This correspondence is focused on channel distortion only and $n(t)$ is discarded to simplify the analysis that follows. If the signals are processed by discrete Fourier transform (DFT) band-pass Mel filters and inside each filter the energy of $x(t)$ and the frequency response of $h(t)$ are considered to be constant, the log-energy of the noisy signal at the output of filter m in frame i can be modeled as [3], [5], [9]

$$\log\left[\overline{y_{i,m}^2}\right] = \log\left[\overline{x_{i,m}^2}\right] + \log\left[\overline{H^2[\omega_m]}\right] \tag{2}$$

where $\overline{x_{i,m}^2}$ and $\overline{y_{i,m}^2}$ are the energy of the clean and distorted speech signals at the output of filter m in frame i, respectively, $H[\omega_m]$ is the frequency response of filter $h(t)$ that models channel distortion, $1 \leq m \leq M$ and M is the number of mel filters; and ω_m is the discrete central frequency of mel filter m. As a result, the observed signal in the MFCC domain can be modeled as [3], [5], [9]

$$Y_{i,n}^C = X_{i,n}^C + H_n^C \tag{3}$$

where $X_{i,n}^C$ and $Y_{i,n}^C$ denote, respectively, static cepstral coefficient n in frame i of $x(t)$ and $y(t)$, $1 \leq n \leq N$, where N is the number of static cepstral coefficients, and H_n^C is the cepstral bias associated with the channel distortion at feature n. Surprisingly, despite the facts that the model in (2) and (3) has widely been employed by many authors and that the channel is modeled as a linear time-invariant filter, the continuity of the frequency response $H(\omega)$ has not been explored exhaustively. In other words, the additive components $\log\left[\overline{H^2[\omega_m]}\right]$ in (2) are usually treated and estimated without considering that $H[\omega_m]$ corresponds to samples of a continuous curve $H(\omega)$.

The feature compensation techniques that make use of the model described in (2) or (3) can be classified as utterance-based or model-based. Utterance-based approaches such as CMN [10], CMVN [11], RASTA [12], MVA [13], and further techniques [14], [15] can dramatically reduce the error rate in channel mismatch conditions, but performance degrades when the amount of testing data is limited [16]. On the other hand, model-based feature compensation methods attempt to estimate the channel bias component shown in (3) with a reference acoustic model by employing maximum-likelihood (ML) or maximum *a posteriori* (MAP) criteria. The most widely adopted model-based approach corresponds to the Gaussian mixture model (GMM) built from reference speech data [1]–[7], where the feature-space correction can be estimated with the expectation–maximization (EM) algorithm. The EM algorithm can lead to significant amelioration of the impact of telephone-channel mismatch conditions but it is computationally costly. Feature-space bias compensation parameters can also be estimated by maximizing the likelihood of testing utterances with a nearest neighbor search, where each observed frame is associated only with its most likely acoustic model unit. The nearest neighbor search can be performed by using two strategies [5]: 1) the most likely Gaussian within a GMM is associated with each frame; and 2) a forced-Viterbi-alignment-based scheme, where each frame is associated with the optimal acoustic Gaussian, state and model in an HMM. Despite the fact that the EM-based methods generally provide higher improvements, the computational load required by nearest neighbor search based estimation with GMM and Viterbi algorithm is substantially lower.

Other approaches such as joint factor analysis (JFA) [17] and nuisance attribute projection (NAP) [18] have been proposed in the context of GMM and/or SVM-based text-independent speaker verification (TI-SV) systems. In those cases, the addressed task considers scenarios where training and testing data correspond to 300 seconds or more of speech samples (e.g., NIST evaluations). As a result, those techniques can hardly be applicable to Viterbi-based text-dependent speaker verification systems with limited data.

In conventional bias removal schemes based on hypotheses H1 and H2 the bias correction components in (2) or (3) are usually estimated independently in each feature dimension. In this context, a relevant

Manuscript received July 14, 2010; revised December 16, 2010; accepted April 10, 2011. Date of publication May 23, 2011; date of current version November 23, 2011. This work was supported by Conicyt-Chile under Grant Fondecyt 1100195. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Haizhou Li.

The authors are with the Speech Processing and Transmission Laboratory, Department of Electrical Engineering, University of Chile, C.P. 412-3, Santiago, Chile (e-mail: cgarreto@ing.uchile.cl, nbecerra@ing.uchile.cl).

Digital Object Identifier 10.1109/TASL.2011.2157495

issue is the number of parameters that are needed to model convolutional noise: the higher the number of model parameters, the greater the amount of required estimation data. Conventional channel feature-space compensation techniques estimate as many parameters as the number of static feature coefficients according to (3). If the number of parameters in the channel distortion model were reduced, a greater improvement in accuracy could be achieved with limited estimation data.

This correspondence proposes a method to improve the accuracy of nearest neighbor-based estimation of channel distortion in text-dependent SV with limited data without increasing the computational load significantly. Given this scenario the most straightforward strategy is to reduce the number of parameters in the channel distortion model. To do so the technique described here models the additive bias in the LFBE domain according to (2) as a polynomial function of m. This is accomplished by imposing the condition that $\log\left[\overline{H^2[\omega_m]}\right]$ are samples of a continuous curve $\log\left[\overline{H^2(\omega)}\right]$, which in turn could be represented as a polynomial function of ω. As mentioned above ω_m is the central frequency of Mel filter m and $\log\left[\overline{H^2[\omega_m]}\right]$ could also be modeled as a polynomial function of m instead of ω_m. Consequently, the additive channel distortion in (2) is fit to a Pth-order polynomial function where $P \leq M$ and M is the number of Mel filters. As a result, channel distortion is also modeled as an additive component in the cepstral domain but, in contrast to (3), the bias component is estimated as a weighted average of the polynomial coefficients. If $P \leq N$, where N is the number of static cepstral coefficients, the number of parameters to estimate is reduced when compared to the ordinary bias-component model as described in (3). As shown here, the polynomial function that models $\log\left[\overline{H^2[\omega_m]}\right]$ in the LFBE domain leads to a linear function in the MFCC domain. Polynomial functions have been successfully employed in acoustic modeling and noise robustness techniques in ASR and SV fields because of their simple form, flexibility of shaping and low computational load [4], [19].

In this correspondence, the proposed polynomial model of channel distortion in the LFBE domain is employed in combination with nearest neighbor search-based estimation with GMM and forced Viterbi alignment. Experiments with telephone speech in limited-data scenarios and channel-mismatch conditions suggest that the presented method can lead to reductions in equal error rate (EER) as great as 22% and 8% when compared with the baseline system and the standard cepstral bias removal approach, respectively. Finally, the polynomial-based channel-distortion cancelation scheme increases the computational load of the verification attempts by just 4.2% when compared with the conventional cepstral bias removal strategy.

II. Polynomial Model of Channel Distortion in the Log Filter-Bank Energy Domain

As mentioned above, the major innovation of the proposed model is to take into consideration the fact that the additive bias components of channel distortion in the LFBE domain according to (2) are samples of a continuous frequency response curve in ω. If an analog telephone line is composed basically of twisted-pair cables and hand-set microphones, it is reasonable to model the telephone channel with a low-pass RC filter that provides a continuous frequency response curve. Clearly, the channel effect on a given spectral component is not independent of the gain introduced in another frequency. Consequently, convolutional distortion could be modeled with a parametric function along the spectrum to reduce the number of parameters to estimate. By doing so, the additive bias in (2) or (3) should be estimated more reliably in limited data scenarios. In order to simplify the formulae $Y_{i,m} = \log\left[\overline{y_{i,m}^2}\right]$, $X_{i,m} = \log\left[\overline{x_{i,m}^2}\right]$ and $G_m = \log\left[\overline{H^2[\omega_m]}\right]$. In this correspondence G_m is modeled as a polynomial function of m

$$G_m = \sum_{p=0}^{P} a_p \cdot m^p \tag{4}$$

where a_p is the pth the polynomial coefficient, P is the polynomial order and $A = \{a_p\}_{p=0}^{P}$. By employing the discrete cosine transform (DCT) and (2), the nth distorted cepstral feature at frame i, $Y_{i,n}^C$, is expressed as

$$\begin{aligned} Y_{i,n}^C &= \sum_{m=1}^{M} Y_{i,m} \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot (m - 0.5)\right) \\ &= \sum_{m=1}^{M} (X_{i,m} + G_m) \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot (m - 0.5)\right). \end{aligned} \tag{5}$$

By replacing G_m with the polynomial approximation in (4), $Y_{n,t}^C$ can be rewritten as

$$\begin{aligned} Y_{i,n}^C = &\sum_{m=1}^{M} \left\{X_{i,m} \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot (m - 0.5)\right)\right\} \\ &+ \sum_{p=0}^{P} \left\{a_p \cdot \sum_{m=1}^{M} m^p \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot (m - 0.5)\right)\right\}. \end{aligned} \tag{6}$$

In a real application $Y_{i,n}^C$ is the observed nth cepstral feature at frame i. By defining

$$W_{p,n} = \sum_{m=0}^{M} m^p \cdot \cos\left(\frac{\pi \cdot n}{M} \cdot (m - 0.5)\right). \tag{7}$$

$Y_{i,n}^C$ in (6) can be expressed as

$$Y_{i,n}^C = X_{i,n}^C + \sum_{p=0}^{P} a_p \cdot W_{p,n}. \tag{8}$$

Notice that $W_{p,n}$ depends only on $0 \leq p \leq P$ and DCT constants $\cos(\pi \cdot n/M \cdot (m - 0.5))$. According to (8) the additive bias in the MFCC domain, $G_n^C(A) = \sum_{p=0}^{P} a_p \cdot W_{p,n}$, is a linear combination of the polynomial coefficients a_p weighted by $W_{p,n}$.

III. Nearest Neighbor-Based Estimation of Channel Distortion Modeled as a Polynomial Function

As explained above, the proposed polynomial channel distortion model is used in combination with nearest neighbor-based algorithms. Consider the observed feature vector sequence, $Y^C = \{Y_i^C\}_{i=0}^{I-1}$, where $Y_i^C = \{Y_{i,n}^C\}_{n=0}^{N-1}$ corresponds to the frame at instant i and I is the number of frames. Frame Y_i^C is associated with one of the acoustic units s_k that belongs to a reference acoustic model λ, where $1 \leq k \leq K$ and K is the number of acoustics units in λ. Consequently, in the case of the GMM-based computation, λ and s_k represent a Gaussian mixture and Gaussian component k, respectively. In estimation based on forced Viterbi alignment, λ and s_k may represent, respectively, a sequence of context-dependent phoneme HMMs, and a Gaussian in a state within this composed HMM. Finally, both GMM and forced-Viterbi-based estimation provide an output denoted by $S = \{s_{k(i)}\}_{i=0}^{I-1}$ that is aligned to feature vector sequence Y^C, where $s_{k(i)}$ denotes the acoustic unit associated with frame Y_i^C. The presented approach involves three main steps.

Step 1) Given a feature vector sequence Y^C, S is obtained employing a nearest neighbor search by means of nn-GMM or forced Viterbi alignment.

Step 2) The feature-space correction is computed by employing the polynomial approximation model according to (4). As a result, the polynomial parameter vector A is estimated.

Step 3) Finally, the compensated frame sequence $\hat{X}^C = \{\hat{X}_i^C\}_{i=0}^{I-1}$ is obtained according to

$$\hat{X}_{i,n}^C = Y_{i,n}^C - \sum_{p=0}^{P} a_p \cdot W_{p,n}. \tag{9}$$

In Step 2, the polynomial function parameter vector A can be estimated by using the ML criterion

$$\hat{A} = \arg\max_A \{p(Y^C|\lambda, S, A)\} \tag{10}$$

where $\hat{A} = \{\hat{a}_p\}_{p=0}^P$ is the optimal parameter vector that defines the polynomial function of (4). The probability density function (pdf) of the acoustic unit s_k is modeled by a Gaussian function with mean vector $\mu_k = \{\mu_{k,n}\}_{n=0}^{N-1}$ and diagonal covariance matrix Σ_k, and $\phi_k = (\mu_k, \Sigma_k)$. The diagonal components of Σ_k are denoted by $\sigma_k^2 = \{\sigma_{k,n}^2\}_{n=0}^{N-1}$. In this case, the likelihood $p(Y_i^C|\phi_{k(i)}, A)$ is defined as

$$p\left(Y_i^C|\phi_{k(i)}, A\right) = \frac{1}{(2\pi)^{N/2}|\Sigma_{k(i)}|^{1/2}} \cdot e^{-1/2\sum_{n=0}^{N-1}\left[Y_{i,n}^C - \left(\sum_{p=0}^P a_p \cdot W_{p,n}\right) - \mu_{k(i),n}\right]^2/\sigma_{k(i),n}^2}. \tag{11}$$

where $\phi_{k(i)} = (\mu_{k(i)}, \Sigma_{k(i)})$ denotes the set of Gaussian parameters associated with component $s_{k(i)}$ allocated to frame Y_i^C. The optimal polynomial coefficient vector $\hat{A}$ can be estimated by maximizing the log-likelihood of the following target function:

$$\begin{aligned}\hat{A} &= \arg\max_A \{\log[p(Y^C|\lambda, S, A)]\} \\ &= \arg\max_A \left\{\sum_{i=0}^{I-1} \log\left[p\left(Y_i^C|\phi_{k(i)}, A\right)\right]\right\}.\end{aligned} \tag{12}$$

By replacing (11) in (12), the optimization can be rewritten as

$$\hat{A} = \arg\max_A \left\{\begin{array}{l}\sum_{i=0}^{I} \log\left[\left((2\pi)^{N/2}|\Sigma_{k(i)}|^{1/2}\right)^{-1}\right] \\ -\frac{1}{2} \cdot \sum_{i=0}^{I-1}\sum_{n=0}^{N-1} \frac{\left[Y_{i,n}^C - \left(\sum_{p=0}^{P} a_p \cdot W_{p,n}\right) - \mu_{k(i),n}\right]^2}{\sigma_{k(i),n}^2}\end{array}\right\} \tag{13}$$

where $\sum_{i=0}^I \log\left[\left((2\pi)^{N/2}|\Sigma_{k(i)}|^{1/2}\right)^{-1}\right]$ does not depend on A and is discarded. As a result, $\hat{A}$ is estimated by computing the partial derivates of (13) with respect to a_q, where $0 \leq q \leq P$, and setting then to zero. Then, the optimization in (13) leads to a linear system of $P+1$ equations and $P+1$ unknown variables:

$$\sum_{p=0}^{P} \hat{a}_p \cdot \left\{\sum_{i=1}^{I}\sum_{n=1}^{N}\left[\frac{(W_{q,n} \cdot W_{p,n})}{\sigma_{k(i),n}^2}\right]\right\} = \sum_{i=1}^{I}\sum_{n=1}^{N}\left[\left(\frac{W_{q,n}}{\sigma_{k(i),n}^2}\right) \cdot \left(Y_{i,n}^C - \mu_{k(i),n}\right)\right]. \tag{14}$$

If $\beta_{q,p} = \sum_{i=1}^I \sum_{n=1}^N (W_{q,n} \cdot W_{p,n})/\sigma_{k(i),n}^2$ and $\gamma_q = \sum_{i=1}^I \sum_{n=1}^N \left(W_{q,n}/\sigma_{k(i),n}^2\right) \cdot \left(Y_{i,n}^C - \mu_{k(i),n}\right)$, the linear system in (14) can be expressed as

$$\sum_{p=1}^{P} \hat{a}_p \cdot \beta_{q,p} = \gamma_q. \tag{15}$$

Let $\Gamma = \{\gamma_q\}_{q=0}^P$ and $B = \{\beta_{q,p}\}_{(P+1)\times(P+1)}$. Consequently, the solution for the system in (15) can be easily rewritten as

$$\hat{A} = \mathrm{B}^{-1} \cdot \Gamma. \tag{16}$$

The parameters Γ and B are estimated for all frames of the utterance as indicated above.

IV. Experiments

The polynomial-based channel distortion compensation scheme proposed in this correspondence was tested with a TD-SV system using a telephone database. The results were obtained with a telephone version of the YOHO database [20], which supports the development, training, and testing of SV systems. The vocabulary is composed of two-digit numbers spoken continuously in sets of three (e.g., "62–31–53"). The database is divided into "enrollment" and "verification" segments; each segment contains data from 138 speakers. In this correspondence a subset of 70 speakers (40 males and 30 females) was employed. The speakers were divided as follows: 40 speakers (20 males and 20 females) to train the speaker-independent (SI) HMM used in the score normalization; and, 30 testing speakers (20 males and 10 females) for verification attempts. For each speaker, one 24-utterance enrollment session was considered. Four verification sessions per testing speaker were employed, with four utterances per session. Each utterance was recorded on a real landline telephone call by employing a speaker/telephone handset acoustic coupling. Seven handsets were used (hset1, hset2,…, hset7). Signals were sampled at 8 kHz and 16 bits per sample. Handset hset1 was labeled as the reference or "matched" channel, and clients' models and SI HMM were generated using the enrollment utterances recorded with hset1. The verification attempts were performed by employing testing utterances recorded with every handset. Consequently, false rejection curves were estimated with 7 handsets × 30 speakers/per handset × 16 verification signals per $\mathrm{client} = 3360$ utterances. False acceptance curves were obtained with 7 handsets × 30 speakers × 29 impostors/per handset × 6 verification signals/per $\mathrm{impostor} = 36540$ experiments. Observe that, as suggested by some authors, the ratio between client and impostor attempts is approximately equal to 1 to 10. Speech signals were divided into 25-ms frames with 50% overlap. The band from 300 to 3400 Hz was covered by 14 Mel DFT filters, and at the output of each channel the logarithm of the energy was computed. The final feature vector at frame i, O_i, is composed of the frame energy plus ten static cepstral coefficients, along with their first and second time derivatives. 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 and eight multivariate Gaussian densities per state in speaker-dependent (SD) and SI models, respectively. The verification score, or the normalized log-likelihood of the feature vector sequence $O = \{O_i\}_{i=0}^I$, $\log L(O)$, is defined as [10]

$$\log L(O) = \log L(O|\lambda_{SD}) - \log L(O|\lambda_{\mathrm{SI}}) \tag{17}$$

where $\log L(O|\lambda_{SD})$ is the log-likelihood of the client hypothesis and λ_{SD} is the SD model associated with the claimed identity, and $\log L(O|\lambda_{\mathrm{SI}})$ is the log-likelihood of the SI model λ_{SI}. In matched-channel conditions (hset1) the equal error rate (EER) given

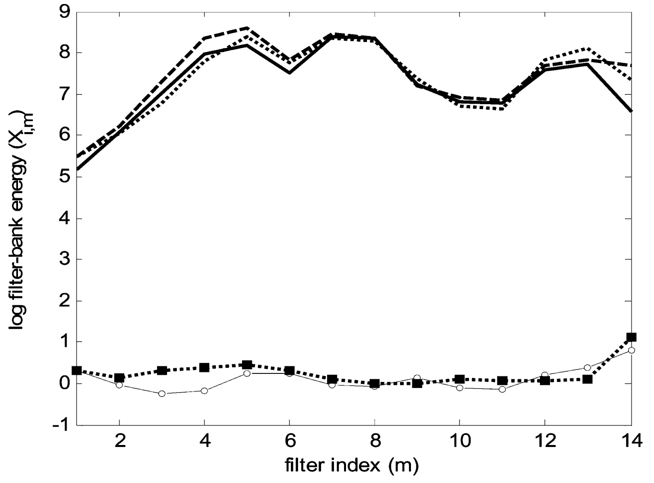

Fig. 1. The curves at the top corresponds to a graphical representation of the log filter-bank energy (LFBE) feature vector of speech frame i recorded with: reference channel hset1 (—); hset2 (- - -); and hset3 (· · ·). The curves at the bottom are the difference in the LFBE domain between: hset2 and reference channel hse1 (· · ■ · ·); and, hset 3 and reference channel hset1 (- o -).

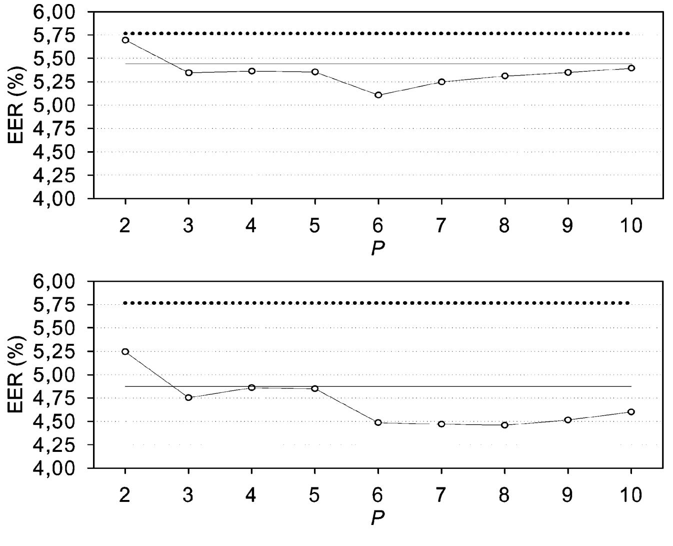

Fig. 2. EER (%) versus polynomial order P as defined in (8): (a) nn-GMM-Poly (- o -), baseline system (· · ·), and nn-GMM-SBR (—); and, (b) Viterbi-Poly (- o -), baseline system (. . .) and Viterbi-SBR (—).

TABLE I
EER(%) Obtained With the Baseline System, RASTA, CMN, and CMVN

	Baseline	RASTA	CMN	CMVN
EER(%)	5.77	5.54	5.51	5.97

by the baseline system is equal to 3.3%. When the whole testing database is used (seven handsets) the baseline EER is equal to 5.77%.

The polynomial model proposed in this correspondence is used in combination with nearest-neighbor GMM estimations and estimations based on forced Viterbi alignment, denoted by nn-GMM-Poly and Viterbi-Poly, respectively. The reference GMM used in nearest neighbor GMM-based estimations was composed of 256 Gaussian mixtures and was generated with the same data employed to train SI HMM λ_{SI} as explained above. In the case of the forced Viterbi alignment, the optimal state alignment is estimated employing λ_{SI} as the reference model. In this case the most likely Gaussian $s_{k(i)}$ is chosen from a set composed of the eight Gaussians in the state from λ_{SI} allocated to Y_i^C plus the Gaussian in the corresponding state within λ_{SD}. The polynomial-model-based channel distortion estimation is compared with a signal bias removal (SBR) strategy [5] that makes use of the ordinary cepstral bias model in (2) and (3). SBR is also employed in combination with nearest-neighbor GMM and forced-Viterbi-alignment-based estimations, nn-GMM-SBR and Viterbi-SBR, respectively.

V. Discussion

Fig. 1 depicts the LFBE feature vector of a given frame recorded with telephone channels hset1 (reference channel), hset2, and hset3. Fig. 1 strongly suggests that the channel distortion, i.e., the difference between LFBE feature vectors, is a continuous curve that can easily be modeled with a polynomial function. Table I shows these results along with RASTA, CMN, and CMVN for comparison purposes. When compared with the baseline system, the reductions in relative EER provided by RASTA and CMN are equal to 4.0% and 4.5%, respectively. Also, Table I shows that the use of CMVN increases the relative error rate by 3.5% when compared with the baseline system. This must be due to inaccurate estimation of the cepstral variance when limited data are available in training and testing, which can lead to inaccurate channel distortion cancellation and a consequent increase of the error rate. According to Fig. 2, the nn-GMM-SBR and Viterbi-SBR procedures provide reductions in relative EER as high as 5.7% and 15.5%, respectively, when compared with the baseline system. The greater improvement provided by Viterbi-SBR when compared with nn-GMM-SBR must be due to the fact that Viterbi-SBR includes temporal information about the sequence of acoustic units and nn-GMM-SBR does not. Also in Fig. 2, Viterbi-Poly provides a reduction in EER as high as 22.7% and 8.4% when compared with the baseline system and Viterbi-SBR, respectively, with $P = 8$. Significance analysis with the McNamar's test [21] shows that these improvements are statistically significant $(p < 0.034)$. In addition, nn-GMM-Poly leads to improvements in EER equal to 11.5% and 6.3% when compared with the baseline system and nn-GMM-SBR, respectively, also with $P = 6$. These results are also statistically significant $(p < 0.01)$. It is interesting to highlight that the polynomial-based channel distortion model leads to higher reductions in EER than the ordinary SBR scheme when $6 \leq P \leq 8$. Notice that the ordinary SBR model requires estimating as many channel distortion components as there are static cepstral coefficients (e.g., ten static MFCC features in this correspondence). Consequently, a reduction in the number of parameters to estimate is clearly achieved.

Figs. 3 and 4 present DET curves provided by the baseline system, the ordinary signal bias model denoted with SBR and the proposed polynomial-based channel distortion model. As can be seen in Figs. 3 and 4, nn-GMM-Poly $(P = 6)$ and Viterbi-Poly $(P = 6)$ give a reduction in the area below the DET curve higher than nn-GMM-SBR and Viterbi-SBR, respectively. This fact shows that both methods reduce EER in the proximity of TEER (threshold of equal error rate) and provide higher discrimination ability. It is worth highlighting that, due to the fact that the polynomial-based channel distortion model is employed with nearest neighbor strategies, the increase in computational load of nn-GMM-Poly and Viterbi-Poly when compared with nn-GMM-SBR and Viterbi-SBR, respectively, in negligible. For instance, the estimation of polynomial parameter vector A requires an average processing time just 18.8% higher than the needed to estimate cepstral bias H^C defined in the ordinary channel distortion model in (3), which in turn represents 22.2% of the total processing time required by the TD-SV engine for one verification attempt. Consequently, the proposed polynomial based compensation scheme increases the processing time required by one verification attempt by just 4.2% when compared with the standard cepstral bias removal approach.

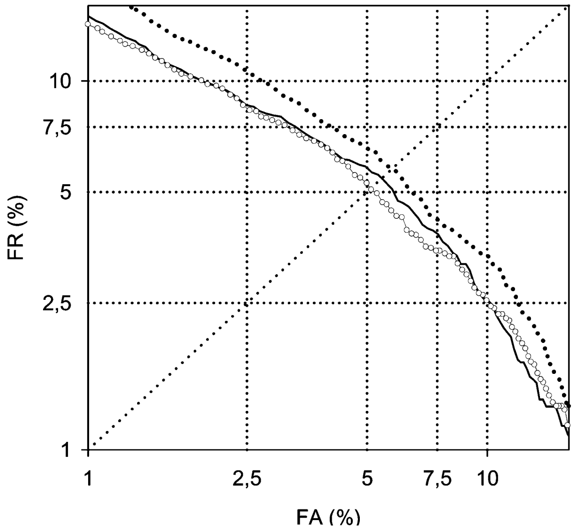

Fig. 3. DET curves obtained with the baseline system $(\cdots)$, nn-GMM-SBR (—) and nn-GMM-Poly, with P defined in (8) equal to 6(- ○ -).

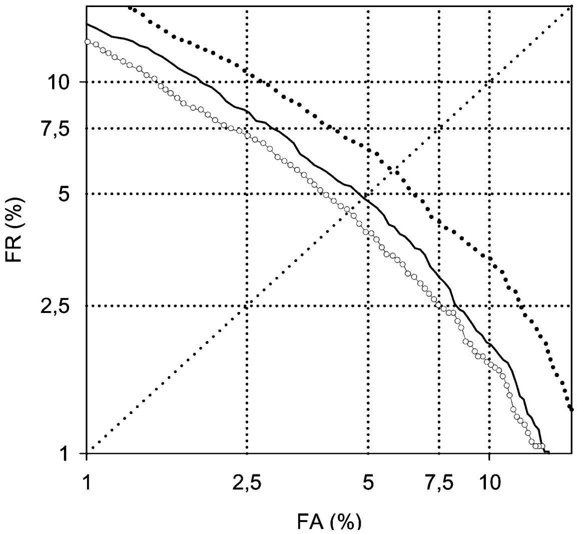

Fig. 4. DET curves obtained with the baseline system $(\cdots)$, Viterbi-SBR (—) and Viterbi-Poly, with P defined in (8) equal to 6(- ○ -).

It is worth noting that the greatest improvement takes place with $P = 6$ and $P = 8$ with nn-GMM and forced-Viterbi-alignment-based estimations, respectively. Despite the fact that the effectiveness of the approach depends on polynomial order P, a significant reduction in EER is achieved when $6 \leq P \leq 8$ with both nn-GMM and forced Viterbi alignment. For instance, although not shown here, Viterbi-Poly with each individual handset in mismatch conditions (hset2, hset3,…, hset7) leads to the lowest EER in four out of six telephone channels when $6 \leq P \leq 8$. In the other two cases, the EER achieved in the interval $6 \leq P \leq 8$ is just 0.7% and 0.3% higher than the lowest EER in the intervals $2 \leq P \leq 5$ or $9 \leq P \leq 10$

Table II shows results with Viterbi-Poly $(6 \leq P \leq 8)$ in combination with RASTA, CMN, and CMVN. As can be seen in Table II, Viterbi-Poly reduces the ERR achieved with RASTA, CMN, and CMVN in 15.2%, 12.9% and 4.2%, respectively. However, the lowest EER is obtained when Viterbi-Poly is applied by itself. This result must be due to the fact that RASTA, CMN, and CMVN tend to lose effectiveness with limited data.

TABLE II
EER(%) OBTAINED WITH VITERBI-POLY WITH $P = \{6, 7, 8\}$ WHEN APPLIED IN ISOLATION AND IN COMBINATION WITH RASTA, CMN, AND CMVN

	Viterbi-Poly $P = 6$	Viterbi-Poly $P = 7$	Viterbi-Poly $P = 8$
Isolated	4.49	4.47	4.46
RASTA	4.60	4.77	4.70
CMN	4.82	4.82	4.80
CMVN	5.98	5.97	5.72

VI. CONCLUSION

A novel polynomial-based feature-space channel compensation method is proposed in this correspondence. The presented technique models channel distortion by employing a polynomial function in the log filter bank energy domain. The method described models the continuity of the channel frequency response and reduces the number of parameters that need to be estimated by imposing appropriate constraints on the channel distortion. Results show that the proposed model can lead to relative reductions in EER as great as 22% and 8%, respectively, when compared with the baseline system and an ordinary cepstral bias removal strategy with limited data. The computational load is kept low by making use of nearest-neighbor GMM and forced-Viterbi-alignment-based estimations, and the polynomial-based scheme increases the processing time by just 4.2% of the whole verification attempt. The application of this technique to other tasks such as speech recognition can be the object of future research.

ACKNOWLEDGMENT

The authors would like to thank Prof. R. Stern from Carnegie Mellon University for having proofread this correspondence.

REFERENCES

[1] M. W. Mak *et al.*, "Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification," in *EURASIP J. Appl. Signal Process.*, 2004, vol. 4, pp. 452–465.

[2] K. K. Yiu *et al.*, "Blind stochastic feature transformation for channel robust speaker verification," *J. VLSI Signal Process.*, vol. 42, no. 2, pp. 117–126, 2006.

[3] J. M. Huerta, "Alignment-based codeword-dependent cepstral normalization," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 451–459, Oct. 2002.

[4] A. Acero *et al.*, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, Beijing, China, 2000, pp. 869–872.

[5] M. G. Rahim and B. H. Juang, "Signal bias removal by maximum-likelihood estimation for robust telephone speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 19–30, Jan. 1996.

[6] H. Jiang *et al.*, "Hierarchical stochastic feature matching for robust speech recognition," in *Proc. ICASSP*, Salt Lake City, UT, 2001, pp. 217–220.

[7] M. Afify *et al.*, "A general joint additive and convolutive bias compensation approach applied to noise lombard speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 6, pp. 524–538, Nov. 1998.

[8] M. J. F. Gales, "Maximum-likelihood linear transformation for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.

[9] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech Recognition," Ph.D. dissertation, Dept. Elect. Comput. Eng., Carnegie Mellon Univ., Pittsburgh, PA, 1990.

[10] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, no. 2, pp. 254–272, Apr. 1981.

[11] R. Zheng *et al.*, "A comparative study of feature and score normalization for speaker verification," *Lecture Notes in Computer Science 3832*, pp. 531–538, 2005.

[12] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[13] C. P. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[14] Z. Tufekci, "Convolutional bias removal based on normalizing the filterbank spectral magnitude," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 485–488, Jul. 2007.

[15] J. W. Hung and L. S. Lee, "Optimization of temporal filters for constructing robust features in speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 808–832, May 2006.

[16] L. Wang *et al.*, "Robust distant speech recognition by combining position-dependent CMN with conventional CMN," *Proc. ICASSP'07*, pp. 817–820, 2007.

[17] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[18] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP'05*, 2005, vol. 1, pp. 629–632.

[19] X. Cui and A. Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1161–1172, Nov. 2005.

[20] J. Campbell and A. Higgins*, YOHO Speaker Verification*. Philadelphia, PA: LDC, 1994.

[21] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proc. ICASSP'89*, 1989, vol. 1, pp. 532–535.