COMBINING STOCHASTIC FEATURE TRANSFORMATION AND HANDSET IDENTIFICATION FOR TELEPHONE-BASED SPEAKER VERIFICATION

Man-Wai Mak

Center for Multimedia Signal Processing
Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University, China

Sun-Yuan Kung[♯]

Dept. of Electrical Engineering
Princeton University
USA

ABSTRACT

The performance of telephone-based speaker verification systems can be severely degraded by the acoustic mismatch caused by telephone handsets. This paper proposes to combine a handset selector with stochastic feature transformation to reduce the mismatch. Specifically, a GMM-based handset selector is trained to identify the most likely handset used by the claimants, and then handset-specific stochastic feature transformations are applied to the distorted feature vectors. To overcome the non-linear distortion introduced by telephone handsets, a 2nd-order stochastic feature transformation is proposed. Estimation algorithms based on the stochastic matching technique and the EM algorithm are derived. Experimental results based on 150 speakers of the HTIMIT corpus show that the handset selector is able to identify the handsets accurately (98.3%), and that both linear and non-linear transformation reduce the error rate significantly (from 12.37% to 5.49%).

1. INTRODUCTION

In recent years, research has focused on verifying speakers' identity over the telephone, primarily because of the recent proliferation of electronic banking and electronic commerce. Telephone-based speaker verification, however, poses a challenge: transducer variability could result in acoustic mismatches between the speech data gathered from different handsets. The sensitivity to handset variations means that handset compensation techniques are essential for practical speaker verification systems.

One possible approach to resolving the mismatch problem is feature transformation.[1] Feature-based approaches attempt to modify the distorted features so that the resulting features fit the clean speech models better. These approaches include cepstral mean subtraction (CMS) [1] and signal bias removal [2], which approximate a linear channel by the long-term average of distorted cepstral vectors. These approaches, however, do not consider the effect of background noise. A more general approach, in which additive noise and convolutive distortion are modeled as codeword-dependent cepstral biases, is the codeword-dependent cepstral normalization (CDCN) [3]. The CDCN, however, only works well when the background noise level is low.

When stereo corpora are available, channel distortion can be estimated directly by comparing the clean feature vectors against their distorted counterparts. For example, in SNR-dependent cepstral normalization (SDCN) [3], cepstral biases for different signal-to-noise ratios are estimated in a maximum likelihood framework. In probabilistic optimum filtering [4], the transformation is a set of multi-dimensional least-squares filters whose outputs are probabilistically combined. These methods, however, rely on the availability of stereo corpora. The requirement of stereo corpora can be avoided by making use of the information embedded in the clean speech models. For example, in stochastic matching [5], the transformation parameters are determined by maximizing the likelihood of observing the distorted features given the clean models.

Although the above methods have been successful in reducing channel mismatches, most of them operate on the assumption that the channel effect can be approximated by a linear filter. Most telephone handsets, in fact, exhibit energy-dependent frequency responses [6] for which a linear filter may be a poor approximation. Recently, this problem has been addressed by considering the distortion as a non-linear mapping [7, 8]. However, these methods rely on the availability of stereo corpora with accurate time alignment.

In this paper, we present a method in which non-linear transformations can be estimated under a maximum likelihood framework, thus eliminating the need for accurately aligned stereo corpora. The only requirement is to record a few utterances uttered by a few speakers using different handsets. These speakers do not need to utter the same set of sentences in the recording sessions, although this may

This work was supported by The Hong Kong Polytechnic University, Grant No. A442 and CERG Grant No. B-Q428. ♯S.Y. Kung was also a Distinguished Chair Professor of The Hong Kong Polytechnic University.

[1] Although model-based transformations can also be applied to channel mismatch compensation, they are not the focus of this paper.

improve the system's performance. This paper also proposes to combine a GMM-based handset selector [9] with stochastic feature transformations to improve the system's practicality. Specifically, each handset is assigned a set of transformation parameters. During verification, the handset selector identifies the most likely handset used by the claimant. The distorted vectors are then transformed according to the transformation parameters of the identified handset.

2. STOCHASTIC FEATURE TRANSFORMATION

Stochastic matching [5] is a popular approach to speaker adaptation and channel compensation. Its main idea is to transform distorted data to fit the clean speech models or to transform the clean speech models to better fit the distorted data. In the case of feature transformation, the channel is represented by either a single cepstral bias (**b**) or a bias together with an affine transformation matrix (A). In the latter case, the component-wise form of the transformed vectors is given by

$$\hat{x}_{t,i} = f_\nu(\mathbf{y}_t)_i = a_i y_{t,i} + b_i \tag{1}$$

where $\mathbf{y}_t$ is a D-dimensional distorted vector, $\nu = \{a_i, b_i\}_{i=1}^D$ is the set of transformation parameters, and f_ν denotes the transformation function. Intuitively, the bias $\mathbf{b}$ compensates the convolutive distortion and the matrix A compensates the effects of noise. The 1st-order transformation in (1), however, has two limitations. First, it assumes that all speech signals are subject to the same degree of distortion, which is certainly incorrect for non-linear channels where signals with higher amplitude are subject to a higher degree of distortion (because of the saturation effect in transducers). Second, the use of a single transformation matrix is inadequate for an acoustic environment with a varying noise level. Here, we propose a new approach to overcome these limitations.

2.1. Non-linear Feature Transformation

Our proposal is based on the notion that different transformation matrices and bias vectors should be applied to transform the vectors in different regions of the feature space. This can be achieved by extending (1) to

$$\hat{x}_{t,i} = f_\nu(\mathbf{y}_t)_i = \sum_{k=1}^{K} g_k(\mathbf{y}_t)(c_{ki} y_{t,i}^2 + a_{ki} y_{t,i} + b_{ki}) \tag{2}$$

where $\nu = \{a_{ki}, b_{ki}, c_{ki}; k = 1, \ldots, K; i = 1, \ldots, D\}$ is the set of transformation parameters and

$$g_k(\mathbf{y}_t) = P(k|\mathbf{y}_t, \Lambda_Y) = \frac{\omega_k^Y p(\mathbf{y}_t|\mu_k^Y, \Sigma_k^Y)}{\sum_{l=1}^K \omega_l^Y p(\mathbf{y}_t|\mu_l^Y, \Sigma_l^Y)} \tag{3}$$

is the posterior probability of selecting the k-th transformation given the distorted speech $\mathbf{y}_t$, the speech model $\Lambda_Y = \{\omega_k^Y, \mu_k^Y, \Sigma_k^Y\}_{k=1}^K$ that characterizes the distorted speech, and the density of the k-th distorted cluster

$$p(\mathbf{y}_t|\mu_k^Y, \Sigma_k^Y) = (2\pi)^{-\frac{D}{2}} |\Sigma_k^Y|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mu_k^Y)^T (\Sigma_k^Y)^{-1} (\mathbf{y}_t - \mu_k^Y)\right\}. \tag{4}$$

Note that when $K = 1$ and $c_{ki} = 0$, (2) is reduced to (1), i.e. the standard stochastic matching is a special case of our proposed approach.

Given a clean speech model $\Lambda_X = \{\omega_j^X, \mu_j^X, \Sigma_j^X\}_{j=1}^K$ derived from the clean speech of several speakers (ten speakers in this work), the maximum likelihood estimates of ν can be obtained by maximizing an auxiliary function

$$\begin{aligned} Q(\nu'|\nu) &= \sum_{t=1}^T \sum_{j=1}^K \sum_{k=1}^K h_j(f_\nu(\mathbf{y}_t)) g_k(\mathbf{y}_t) \\ &\quad \cdot \log\left\{\omega_j^X \omega_k^Y p(\mathbf{y}_t|\mu_j^X, \Sigma_j^X, \nu_k')\right\} \\ &= \sum_{t=1}^T \sum_{j=1}^K \sum_{k=1}^K h_j(f_\nu(\mathbf{y}_t)) g_k(\mathbf{y}_t) \\ &\quad \cdot \log\left\{\omega_j^X \omega_k^Y p(f_{\nu_k'}(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X) \cdot |J_{\nu_k'}(\mathbf{y}_t)|\right\} \end{aligned} \tag{5}$$

with respect to ν'. In (5) ν' and ν represent respectively the new and current estimates of the transformation parameters, T is the number of distorted vectors, $\nu_k' = \{a_{ki}', b_{ki}', c_{ki}'\}_{i=1}^D$ denotes the k-th transformation, $|J_{\nu_k'}(\mathbf{y}_t)|$ is the determinant of the Jacobian matrix whose (r, s)-th entry is given by $J_{\nu_k'}(\mathbf{y}_t)_{rs} = \partial f_{\nu_k'}(\mathbf{y}_t)_s / \partial y_{t,r}$, and $h_j(f_\nu(\mathbf{y}_t))$ is the posterior probability given by

$$\begin{aligned} h_j(f_\nu(\mathbf{y}_t)) &= P(j|\Lambda_X, \mathbf{y}_t, \nu) \\ &= \frac{\omega_j^X p(f_\nu(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X)}{\sum_{l=1}^K \omega_l^X p(f_\nu(\mathbf{y}_t)|\mu_l^X, \Sigma_l^X)} \end{aligned} \tag{6}$$

where

$$p(f_\nu(\mathbf{y}_t)|\mu_j^X, \Sigma_j^X) = (2\pi)^{-\frac{D}{2}} |\Sigma_j^X|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2}(f_\nu(\mathbf{y}_t) - \mu_j^X)^T (\Sigma_j^X)^{-1} (f_\nu(\mathbf{y}_t) - \mu_j^X)\right\}. \tag{7}$$

Ignoring the terms independent of ν' and assuming diagonal covariance (i.e. $\Sigma_j^X = \text{diag}\left\{(\sigma_{j1}^X)^2, \ldots, (\sigma_{jD}^X)^2\right\}$ and likewise for Σ_k^Y), (5) can be written as (8) as shown on the top of next page. The generalized EM algorithm can be applied to find the maximum likelihood estimates of ν. Specifically, in the E-step, we use (6) and (7) to compute $h_j(f_\nu(\mathbf{y}_t))$; then in the M-step, we update ν' according to

$$\nu' \leftarrow \nu' + \eta \partial Q(\nu'|\nu)/\partial \nu' \tag{9}$$

where η ($= 0.001$ in this work) is a positive learning factor.[2] These E- and M-steps are repeated until $Q(\nu'|\nu)$ ceases to increase.

[2] In this work, (9) was repeated 20 times in each M-step.

$$Q(\nu'|\nu) = \sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{k=1}^{K} h_j(f_\nu(\mathbf{y}_t))g_k(\mathbf{y}_t)\left\{-\frac{1}{2}\sum_{i=1}^{D}\frac{(c'_{ki}y_{t,i}^2 + a'_{ki}y_{t,i} + b'_{ki} - \mu_{ji}^X)^2}{\left(\sigma_{ji}^X\right)^2} + \sum_{i=1}^{D}\log(2c'_{ki}y_{t,i} + a'_{ki})\right\} \quad (8)$$

The posterior probabilities $g_k(\mathbf{y}_t)$ and $h_j(f_\nu(\mathbf{y}_t))$ suggest that there are K regions in the distorted feature space and K regions in the clean feature space. As a result, there are K^2 possible transformations. We can, however, reduce this number to K by arranging the indexes j and k such that the symmetric divergence

$$D(\Lambda_{X,j}||\Lambda_{Y,k}) = \frac{1}{2}tr\left\{(\Sigma_j^X)^{-1}\Sigma_k^Y + (\Sigma_k^Y)^{-1}\Sigma_j^X - 2I\right\} + \frac{1}{2}(\mu_j^X - \mu_k^Y)^T\left[(\Sigma_k^Y)^{-1} + (\Sigma_j^X)^{-1}\right](\mu_j^X - \mu_k^Y)$$

between the j-th mixture of Λ_X and the k-th mixture of Λ_Y is minimal.

2.2. Piece-wise Linear Feature Transformation

When $c_{ji} = 0$, (2) becomes a piece-wise linear version of the standard stochastic matching (1). The maximum likelihood estimate of ν can be obtained by the EM algorithm. Specifically, in the M-step, we set the derivative of the Q-function in (8) with respect to ν' to zero, which results in

$$s\cdot(a'_{ki})^2 + gb'_{ki} - u - v = 0 \quad \text{and} \quad qa'_{ki} + rb'_{ki} - p = 0$$

where

$$p = \sum_{t=1}^{T}\sum_{j=1}^{K} h_{tj}g_{tk}\mu_{ji}^X/(\sigma_{ji}^X)^2,$$
$$q = \sum_{t=1}^{T}\sum_{j=1}^{K} h_{tj}g_{tk}y_{t,i}/(\sigma_{ji}^X)^2,$$
$$r = \sum_{t=1}^{T}\sum_{j=1}^{K} h_{tj}g_{tk}/(\sigma_{ji}^X)^2,$$
$$s = \sum_{t=1}^{T}\sum_{j=1}^{K} h_{tj}g_{tk}y_{t,i}^2/(\sigma_{ji}^X)^2,$$
$$u = \sum_{t=1}^{T}\sum_{j=1}^{K} h_{tj}g_{tk}\mu_{ji}^X y_{t,i}/(\sigma_{ji}^X)^2, \text{ and}$$
$$v = \sum_{t=1}^{T}\sum_{j=1}^{K} h_{tj}g_{tk}$$

where $h_{tj} = h_j(f_\nu(\mathbf{y}_t))$ and $g_{tk} = g_k(\mathbf{y}_t)$ are estimated during the E-step.

3. HANDSET SELECTOR

Unlike speaker adaptation where the transformation parameters can be estimated during recognition, in speaker verification we need to estimate the transformation parameters before verification takes place. This is because we do not know the claimant's identity in advance. If the transformation parameters are estimated based on claimant's speech obtained in a single verification session only, all the transformed vectors, regardless of the claimant's genuineness, will be mapped to a region very close to the claimed model in the clean feature space. As a result, the claimant will likely be accepted regardless of whether he/she is a genuine speaker or an impostor.

Therefore, to apply stochastic transformation to telephone-based speaker verification, we need to derive one set of transformation parameters for each type of handsets that the users are likely to use. During verification, the transformation parameters corresponding to the most likely handset are used to transform the distorted features. This can be achieved by applying our recently proposed handset selector [9]. Specifically, each handset is associated with one set of transformation parameters; during verification, an utterance of claimant's speech is fed to H GMMs (denoted as $\{\Gamma_k\}_{k=1}^H$). The most likely handset is selected according to

$$k^* = \arg\max_{k=1}^{H}\sum_{t=1}^{T}\log p(\mathbf{y}_t|\Gamma_k) \quad (10)$$

where $p(\mathbf{y}_t|\Gamma_k)$ is the likelihood of the k-th handset. Then, the transformation parameters corresponding to the k^*-th handset are used to transform the distorted vectors.

4. EXPERIMENTS AND RESULTS

The HTIMIT corpus [10] was used to evaluate the proposed approaches. HTIMIT was obtained by playing back a subset of the TIMIT corpus through 9 different telephone handsets and one Sennheizer head-mounted microphone. It is particularly appropriate for studying telephone transducer effects.

Speakers in the corpus were divided into a speaker set (50 male and 50 female) and an impostor set (25 male and 25 female). Each speaker was assigned a personalized 32-center GMM that models the characteristics of his/her own voice. For each GMM, the feature vectors derived from the SA and SX sentence sets of the corresponding speaker were used for training. A collection of all SA and SX sentences uttered by all speakers in the speaker set was used to train a 64-center GMM background model ($\mathcal{M}_b$). The feature vectors were 12-th order LP-derived cepstral coefficients computed at a frame rate of 14 ms using a Hamming window of 28 ms.

For each handset in the corpus, the SA and SX sentences of 10 speakers were used to create a 2-center GMM.[3] For each handset, a set of feature transformation parameters

[3]Only a few speakers will be sufficient for creating these models. However, we did not attempt to determine the optimum number.

Row	Trans. Method	Equal Error Rate (%)										
		cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average	senh
1	Baseline	7.89	6.93	26.96	18.53	5.79	14.09	7.80	13.85	9.51	12.37	2.98
2	CMS	5.81	5.02	12.07	9.41	5.26	8.88	8.44	6.90	6.97	7.64	3.58
3	ST0, $K = 1$	4.06	3.63	8.86	6.05	3.57	6.78	6.66	4.79	5.43	5.54	2.99
4	ST0, $K = 2$	4.27	3.74	9.19	6.74	3.68	6.95	7.06	5.00	5.38	5.78	3.09
5	ST1, $K = 1$	4.33	4.06	8.92	6.26	4.30	7.44	6.39	4.83	6.32	5.87	3.47
6	ST1, $K = 2$	4.27	3.84	9.14	6.73	3.83	7.01	6.98	5.04	5.74	5.84	3.16
7	ST2, $K = 1$	4.10	3.65	8.98	6.06	3.63	6.94	7.23	4.87	5.41	**5.65**	**3.03**
8	ST2, $K = 2$	4.04	3.57	8.85	6.82	3.53	6.43	6.41	4.76	5.02	**5.49**	**2.98**

Table 1. Equal error rates (in %) achieved by the baseline, cepstral mean subtraction (CMS), and different transformation approaches. ST0, ST1, and ST2 stand for stochastic transformation with 0th- 1st-, and 2nd-order, respectively. The enrollment handset is "senh". The last column represents the case where enrollment and verification use the same handset. The average handset identification accuracy is 98.29%. Note that the baseline and CMS do not require the handset selector.

ν were computed based on the estimation algorithms described in Section 2. Specifically, the utterances from handset "senh" were used to create Λ_X, while those from other 9 handsets were used to create $\Lambda_{Y_1}, \ldots, \Lambda_{Y_9}$. The number of transformations for all handset was set to 2 (i.e. $K = 2$).

During verification, a vector sequence $\mathbf{Y}$ derived from a claimant's utterance (SI sentence) was fed to the GMM-based handset selector $\{\Gamma_i\}_{i=1}^{10}$. A set of transformation parameters was selected according to the handset selector's outputs (10). The features were transformed and then fed to a 32-center GMM speaker model ($\mathcal{M}_s$) to obtain a score ($\log p(\mathbf{Y}|\mathcal{M}_s)$), which was then normalized according to

$$S(\mathbf{Y}) = \log p(\mathbf{Y}|\mathcal{M}_s) - \log p(\mathbf{Y}|\mathcal{M}_b)$$

where $\mathcal{M}_b$ is a 64-center GMM background model. $S(\mathbf{Y})$ was compared with a threshold to make a verification decision. In this work, the threshold for each speaker was adjusted to determine the equal error rate (EER). Similar to [11], the vector sequence was divided into overlapping segments to increase the resolution of the error rates.

Table 1 compares different stochastic feature transformation approaches against cepstral mean subtraction (CMS) and the baseline (without any compensation). All error rates were based on the average of 100 genuine speakers and 50 impostors. Evidently, all cases of stochastic feature transformation show significant reduction in error rates. In particular, 2nd-order stochastic transformation achieves the highest reduction. However, the difference in error rates among various stochastic transformations is not significant, which suggests that zero-th order transformation may already be sufficient for systems with limited computation power.

The last column of Table 1 shows that when the enrollment and verification sessions use the same handset (senh), CMS can degrade the performance. On the other hand, in the case of feature transformation (Rows 3 to 8), the handset selector is able to detect the fact that the claimants use the enrollment handset. As a result, the error rates become very close to the baseline. This suggests that the combination of handset selector and stochastic transformation can maintain the performance under matched conditions.

5. CONCLUSIONS

We have presented a new channel compensation approach to telephone-based speaker verification. Results based on 150 speakers of HTIMIT show that combining stochastic transformation with handset identification can significantly reduce verification error rate. Results also demonstrate that linear and non-linear stochastic transformation attain a comparable amount of error reduction, with the non-linear one achieving a slightly better result.

6. REFERENCES

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

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

[3] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Pub., Dordrecht, 1992.

[4] L. Neumeyer and M. Weintraub, "Probabilistic optimal filtering for robust speech recognition," in *Proc. ICASSP'94*, 1994, pp. 417–420.

[5] A. Sankar and C. H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.

[6] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, and G. C. O'Leary, "The effects of telephone transmission degradations on speaker recognition performance," in *ICASSP95*, 1995, pp. 329–332.

[7] X. Li, M. W. Mak, and S. Y. Kung, "Robust speaker verification over the telephone by feature recuperation," in *Proc. Int. Sym. on Intelligent Multimedia, Video and Speech Processing*, 2001, pp. 433–436.

[8] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5, pp. 567–584, 2000.

[9] K. K. Yiu, M. W. Mak, and S. Y. Kung, "A GMM-based handset selector for channel mismatch compensation with applications to speaker identification," in *2nd IEEE Pacific-Rim Conference on Multimedia*, 2001.

[10] D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in *ICASSP'97*, 1997, vol. 2, pp. 1535–1538.

[11] M. W. Mak and S. Y. Kung, "Estimation of elliptical basis function parameters by the EM algorithms with application to speaker verification," *IEEE Trans. on Neural Networks*, vol. 11, no. 4, 2000.