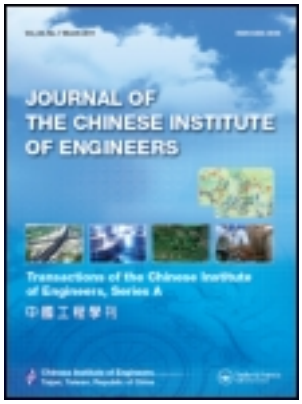


Journal of the Chinese Institute of Engineers
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tcie20

Soft-decision a priori knowledge interpolation for robust telephone speaker identification

Yuan-Fu Liao [a] , Jyh-Her Yang [b] & Sin-Horng Chen [b]

[a] Department of Electronic Engineering , National Taipei University of Technology , Taipei 106, Taiwan, R.O.C. Phone: 886-919-968592 Fax: 886-919-968592 E-mail:

[b] Department of Communication Engineering , National Chiao Tung University , Hsinchu 300, Taiwan, R.O.C.

Published online: 04 Mar 2011.

To cite this article: Yuan-Fu Liao , Jyh-Her Yang & Sin-Horng Chen (2009) Soft-decision a priori knowledge interpolation for robust telephone speaker identification, Journal of the Chinese Institute of Engineers, 32:5, 627-637, DOI: 10.1080/02533839.2009.9671545

To link to this article: http://dx.doi.org/10.1080/02533839.2009.9671545

SOFT-DECISION A PRIORI KNOWLEDGE INTERPOLATION FOR ROBUST TELEPHONE SPEAKER IDENTIFICATION

Yuan-Fu Liao*, Jyh-Her Yang, and Sin-Horng Chen

ABSTRACT

Handsets which are not seen in the training phase (a.k.a unseen handsets) are main sources of performance degradation for speaker identification (SID) applications in telecommunication environments. To alleviate the problem, a soft-decision *a priori* knowledge interpolation (SD-AKI) method of handset characteristic estimation for handset mismatch-compensated SID is proposed in this paper. The idea of the SD-AKI method is to first collect a set of characteristics of seen handsets in the training phase, and to then estimate the characteristic of the unknown testing handset by interpolating the set of seen handset characteristics in the test phase. The estimated handset characteristic is then used to compensate for handset mismatch for robust SID. The SD-AKI method can be realized in both feature and model spaces. Experimental results on the handset TIMIT (HTIMIT) database showed that both the proposed feature- and model-space SD-AKI schemes were more robust than the blind cepstral mean subtraction (CMS), feature warping (FW) methods and their hard-decision counterpart (HD-AKI) for both cases of all-handset and unseen-handset SID tests. It is therefore a promising robust SID method.

Key Words: robust speaker identification, channel mismatch compensation, speech processing.

I. INTRODUCTION

In this paper, the problem of robust speaker identification (SID) (Campbell, 1997; Chaudhari *et al.*, 2003; Faundez-Zanuy and Monte-Moreno., 2005) in the public telephone switching network (PTSN) is addressed. An SID system in PTSN needs to be robust against distortions of various handsets. The characteristic mismatches between the training handsets, where speakers have registered their voiceprints, and testing handsets may cause significant performance degradation (Reynolds and Rose, 1995; Mamone *et al.*, 1996; Murthy *et al.*, 1999). This is especially true for testing handsets whose characteristics are not seen in the training phase (a.k.a unseen handsets). The problem of handset mismatch compensation is not easily solved because the characteristics of the handset and speaker are usually tightly mixed or coupled together. To separate them is essentially a difficult one-to-many mapping problem, unless some *a priori* knowledge about the unknown testing handset is available.

The topic of handset mismatch compensation for speech recognition (Bates and Ostendorf, 1999; Chien and Wang 1998; Gong, 2005; Hermansky and Morgan, 1994; Jiang and Deng, 2000; Juang and Rahim, 1996; Kozat *et al.*, 2007; Sankar and Lee, 1996; Zhao, 2000) has been investigated for many years and various methods have been developed in the past. Recently, some of the proposed approaches have been applied to robust speaker recognition. They can usually be classified into three categories: (1) robust feature-based method, (2) blind feature compensation-based method, and (3) hard-decision *a priori* knowledge-based method.

In the robust feature-based method, features which are less sensitive to handset/channel mismatch, such as pitch dynamics, prosodic and lexical features

*Corresponding author. (Tel: 886-919-968592; Fax: 886-2-27317120; Email: yfliao@ntut.edu.tw)

Y. F. Liao is with the Department of Electronic Engineering, National Taipei University of Technology, Taipei 106, Taiwan, R.O.C.

J. H. Yang and S. H. Chen are with the Department of Communication Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, R.O.C.

(Adami *et al.*, 2003; Chen *et al.*, 2005), Hurst parameter (Sant'Ana *et al.*, 2006), and autocorrelation-based features (You *et al.*, 2005), are extracted and used as auxiliary features to assist the conventional cepstral features in improving the SID recognition performance. In the blind feature compensation-based method, feature processing techniques, such as cepstral mean subtraction (CMS) (Furui, 1981), the RalAtive SpecTrAl (RASTA) method (Hermansky and Morgan, 1994), principal component analysis (PCA)-based temporal filtering (Huan and Lee, 2006), feature warping (FW) (Pelecanos and Sridhari, 2001) and short-time Gaussianization (SG) (Xiang *et al.*, 2002), are applied to blindly compensate/eliminate the handset effect. In the hard-decision *a priori* knowledge-based method, the handset type, such as carbon button or electret, of the testing speech is first detected. Then, a feature-/model-space compensation or transformation technique, such as feature transformation (FT) (Quatieri *et al.*, 2000; Yiu *et al.*, 2004), feature mapping (FM) (Mason *et al.*, 2005; Reynolds, 2003), or speaker model synthesis (SMS) (Beaufays and Weintraub, 1997; Teunen *et al.*, 2000), is then applied to remove the handset distortion or to adapt/synthesize the speaker models using the *a priori* knowledge of the characteristic of the handset type.

Although all those methods are shown to be effective in improving the performance of SID, they may have the following drawbacks. The CMS, RASTA and FW methods may remove not only the characteristic of the handset but also the speaker's as well. The hard-decision *a priori* knowledge-based methods may have difficulties in dealing with the testing speech from an unseen handset since the knowledge about the distortion of the unseen handset is not available. In such a case, one may select the most likely handset from a set of seen handsets, simply reject the testing speech as from an out-of-handset (OOH) source (Mak and Kung, 2002), or backoff to a blind feature compensation-based method.

To alleviate the problem of unseen handsets, a soft-decision *a priori* knowledge interpolation (SD-AKI) method (Yang and Liao, 2004) of handset characteristic estimation for robust SID is proposed in this paper. The idea of the SD-AKI method is to collect a set of characteristics of seen handsets in the training phase and take it as the *a priori* knowledge to represent the space of handsets. In the test phase, the characteristic of the unknown testing handset is first estimated by a soft-decision interpolation using the set of *a priori* handset characteristics, and then used to compensate for the handset mismatch in the SID test.

This paper is organized as follows. Section II presents the proposed SD-AKI handset mismatch-compensated SID method, especially, how to determine the interpolation weights. Experimental results on the HTIMIT database (Reynolds, 1997) are reported in Section III. Some conclusions are given in the last section.

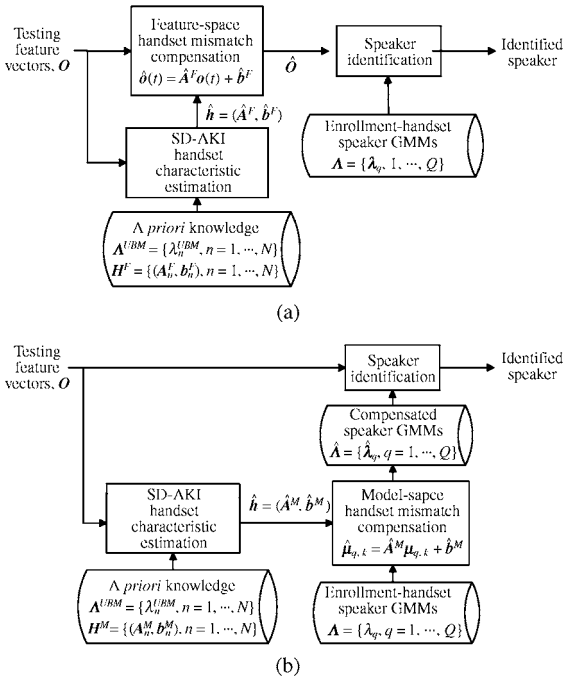


Fig. 1 Block diagrams of the two proposed SD-AKI handset mismatch-compensated speaker identification schemes: (a) feature-space and (b) model-space schemes

II. THE PROPOSED SD-AKI METHOD

The proposed SD-AKI handset mismatch-compensated SID method is to first estimate the characteristics of an unknown testing handset by

$$\hat{\boldsymbol{h}} = \sum_{n=1}^{N} \alpha_n h_n , \tag{1}$$

and to then use $\hat{\boldsymbol{h}}$ to compensate for the handset mismatch for robust SID. Here $\boldsymbol{H} = \{h_n, n = 1, \cdots, N\}$ is a set of *a priori* handset characteristics collected from N seen handsets and $\boldsymbol{\alpha} = \{\alpha_n, n = 1, \cdots, N\}$ are the interpolation weight vector.

Two types of h_n are considered in this study. One is the feature-space transformation function between the n-th seen handset and a testing handset where speakers have registered their voiceprints. This type of h_n can be extracted by the stochastic matching (SM) method (Sankar and Lee, 1996) or the feature maximum likelihood linear regression (FMLLR) method (Gales 1997; Kozat *et al.*, 2007). Another is the model-space transformation function between the n-th seen handset and the testing handset, and can be calculated by the MLLR or constrained MLLR (CMLLR) (Digalakis *et al.*, 1995; Gales and

Woodland, 1996) method. So, the proposed SD-AKI handset mismatch-compensated SID method can be realized in either feature or model space.

Figures 1(a) and 1(b) show block diagrams of the two proposed feature- and model-space SD-AKI handset mismatch-compensated SID schemes, respectively. As shown in these two figures, both schemes prepare *a priori* knowledge of the characteristics of all seen handsets, $\boldsymbol{H}$, and a set of handset-specific universal background models (UBMs), $\boldsymbol{\Lambda}^{UBM} = \{\lambda_n^{UBM},\ n = 1, \cdots, N\}$, in advance in the training phase. In the test phase, they first estimate the handset characteristic, $\hat{\boldsymbol{h}}$, of the input testing speech, and then use it to transform the input speech feature vectors, $\boldsymbol{O}$, to handset mismatch-compensated feature vectors, $\hat{\boldsymbol{O}}$, to match with the characteristic of testing-handset speaker Gaussian mixture models (GMMs), $\boldsymbol{\Lambda} = \{\lambda_q,\ q = 1, \cdots, Q\}$, for SID in the feature-space scheme; or to transform the set of speaker GMMs, $\boldsymbol{\Lambda}$, to a set of testing-handset speaker GMMs, $\hat{\boldsymbol{\Lambda}} = \{\lambda_q,\ \lambda_q = 1, \cdots, Q\}$, to match with the characteristics of the input speech for SID in the model-space scheme.

In the following subsections, we discuss these two schemes in detail:

1. The *A priori* Knowledge of Handset Characteristics

Some well-known speaker adaptation methods MLLR and CMLLR (Sankar and Lee, 1996; Gales 1997; Kozat *et al.*, 2007; Digalakis *et al.*, 1995; Gales and Woodland, 1996), can be used to extract the characteristics of seen handsets from their training data. In this study, the FMLLR and MLLR methods are employed to extract these two types of model- and feature-space *a priori* handset characteristics.

In both schemes, we first train an UBM (Reynolds *et al.*, 2000), $\boldsymbol{\lambda}^{UBM} = \{(c_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),\ k = 1, \cdots, K\}$, in advance using the observed spectral feature vectors of all speakers from the training handset. Here, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k-th mixture component; c_k is the mixture weight; and K is the number of mixture components.

We then use MLLR to train a set of handset-specific UBMs, $\boldsymbol{\Lambda}_n^{UBM} = \{\boldsymbol{\lambda}_n^{UBM},\ n = 1, \cdots, N\}$, where $\boldsymbol{\lambda}_n^{UBM} = \{(c_k, \boldsymbol{\mu}_{n,k}, \boldsymbol{\Sigma}_k),\ k = 1, \cdots, K\}$, via finding the optimal transformation from the UBM, $\boldsymbol{\lambda}^{UBM}$, to $\boldsymbol{\lambda}_n^{UBM}$ using the training feature vectors of all speakers from the n-th seen handset, $\boldsymbol{O}_n = \{\boldsymbol{o}_n(1), \cdots, \boldsymbol{o}_n(T_n)\}$. The relationship between $\boldsymbol{\lambda}_n^{UBM}$ and $\boldsymbol{\lambda}^{UBM}$ can be expressed by

$$\boldsymbol{\mu}_{n,k} = \boldsymbol{A}_n^M \cdot \boldsymbol{\mu}_k + \boldsymbol{b}_n^M, \quad \text{for } k = 1, \cdots, K \text{ and } n = 1, \cdots, N, \tag{2}$$

where $\boldsymbol{A}_n^M$ and $\boldsymbol{b}_n^M$ are the mixture-mean transformation matrix and the bias vector, respectively. Here, only the means of all mixture components are treated as handset-specific because they are the most important factors for mismatch-compensated speaker identification (Gales and Woodland, 1996; Reynolds *et al.* 2000). We regard $\boldsymbol{A}_n^M$ and $\boldsymbol{b}_n^M$ as the characteristics of the n-th seen handset and collect them to form the model-space *a priori* knowledge, i.e., $\boldsymbol{H}^M = \{(\boldsymbol{A}_n^M, \boldsymbol{b}_n^M),\ n = 1, \cdots, N\}$.

For the feature-space scheme, we want to find the optimal transformation of the observed spectral feature vectors, $\boldsymbol{O}_n = \{o_n(1), \cdots, o_n(T_n)\}$, of the n-th seen handset to the space of the UBM by using the following linear regression function

$$\hat{\boldsymbol{o}}(t) = \boldsymbol{A}_n^F \boldsymbol{o}_n(t) + \boldsymbol{b}_n^F, \quad \text{for } t = 1, \cdots, T_n \text{ and } n = 1, \cdots, N, \tag{3}$$

where $\boldsymbol{A}_n^F$ and $\boldsymbol{b}_n^F$ are the transformation matrix and the bias vector, respectively. Here, we use the FMLLR method (Gales 1997; Kozat *et al.*, 2007) to find $\boldsymbol{A}_n^F$ and $\boldsymbol{b}_n^F$. We regard $\boldsymbol{A}_n^F$ and $\boldsymbol{b}_n^F$ as the characteristic of the n-th seen handset to form the feature-space *a priori* knowledge, i.e., $\boldsymbol{H}^F = \{(\boldsymbol{A}_n^F, \boldsymbol{b}_n^F),\ n = 1, \cdots, N\}$.

2. Estimation of the Testing Handset Characteristic

The SD-AKI handset characteristic estimation is essentially a modified *a posteriori*-weighting interpolation method with weights steered by a divergence-based distances measure (between the characteristics of the unknown test handset and the set of seen reference handsets).

The block diagram of SD-AKI is displayed in Fig. 2. It first calculates the log-likelihood vector, $\boldsymbol{L} = \{L(\boldsymbol{O}|\boldsymbol{\lambda}_n^{UBM}),\ n = 1, \cdots, N\}$, of the unknown testing handset with respect to the N seen handsets by the Likelihood Estimator using the feature vectors, $\boldsymbol{O}$, of the input testing speech and the set of handset-specific UBMs, $\boldsymbol{\Lambda}^{UBM}$. These N log-likelihoods are then converted to the *a posteriori* probability vector, $\boldsymbol{P} = \{p(\boldsymbol{\lambda}_n^{UBM}|\boldsymbol{O}),\ n = 1, \cdots, N\}$, by the *A Posteriori* Estimator.

A Jensen-Shannon divergence measure, $D(\boldsymbol{P}, \boldsymbol{U})$ (sometimes called Jensen difference, Burbea and Rao 1982; Schütze and Manning, 1999), is then computed by the Divergence Measurer using the *a posteriori* probability vector, $\boldsymbol{P}$, and a uniformly-distributed reference probability mass function, $\boldsymbol{U} = \{\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N}\}$. The aim is to measure how close the characteristics of the unknown test handset are to the set of the pre-prepared seen handsets. If $\boldsymbol{P}$ is far away from

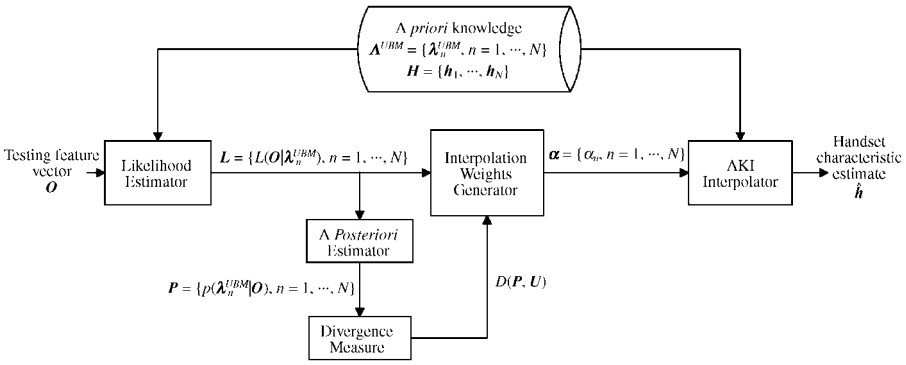


Fig. 2 A block diagram of the SD-AKI method for characteristic estimation of unknown testing handset

$\boldsymbol{U}$, then $D(\boldsymbol{P}, \boldsymbol{U})$ approaches a large value to indicate that the handset characteristic of the testing speech resembles one handset or a few seen handsets. We should therefore emphasize the contributions of those similar handsets to synthesize the unknown handset characteristic. On the contrary, if $\boldsymbol{P}$ is close to $\boldsymbol{U}$, then $D(\boldsymbol{P}, \boldsymbol{U})$ approaches zero. This indicates that the handset characteristic of the testing speech does not fit any of the N seen handsets. We could at best approximate the unknown-handset characteristic using the average characteristic of all seen handsets.

The interpolation weight vector, $\boldsymbol{\alpha}$, are then calculated by the Interpolation Weight Generator using both $\boldsymbol{L}$ and $D(\boldsymbol{P}, \boldsymbol{U})$. Lastly, the testing handset characteristic estimate, $\hat{\boldsymbol{h}}$, is calculated by the AKI Interpolator. We discuss the SD-AKI handset characteristic estimation in detail as follows.

Assume that there are N seen handsets that speakers are likely to use. In the training phase, we first extract the *a priori* handset characteristics, $\boldsymbol{H}$, and derive the set of handset-specific UBMs, $\boldsymbol{\Lambda}^{UBM}$, from the training speech $\boldsymbol{O}_n$ of all N seen handsets. We then take $\boldsymbol{H}$ and $\boldsymbol{\Lambda}^{UBM}$ to represent the space of handsets.

In the test phase, the spectral feature vectors, $\boldsymbol{O}$, of the input speech of a testing speaker from an unknown handset are first fed into the Likelihood Estimator to compute N log likelihoods by

$$L(\boldsymbol{O}|\boldsymbol{\lambda}_n^{UBM}) = \sum_{t=1}^{T} \log p(\boldsymbol{o}(t)|\boldsymbol{\lambda}_n^{UBM}), \quad \text{for } n = 1, \cdots, N, \tag{4}$$

They are then transformed into N *a posteriori* probabilities by

$$p(\boldsymbol{\lambda}_n^{UBM}|\boldsymbol{O}) = \frac{\exp(L(\boldsymbol{O}|\boldsymbol{\lambda}_n^{UBM}))}{\sum_{j=1}^{N} \exp(L(\boldsymbol{O}|\boldsymbol{\lambda}_j^{UBM}))}, \quad \text{for } n = 1, \cdots, N, \tag{5}$$

The *a posteriori* probability vector $\boldsymbol{P}$ made of these N probabilities is then used to calculate the divergence measure, $D(\boldsymbol{P}, \boldsymbol{U})$ (Burbea and Rao 1982; Schütze and Manning, 1999) to evaluate the reliability of applying the *a priori* knowledge to the input testing speech. The divergence measure is defined based on comparing the *a posteriori* probability vector, $\boldsymbol{P}$, with $\boldsymbol{U}$, and can be expressed by

$$D(\boldsymbol{P}, \boldsymbol{U}) = S(\frac{\boldsymbol{P} + \boldsymbol{U}}{2}) - \frac{1}{2}[S(\boldsymbol{P}) + S(\boldsymbol{U})], \tag{6}$$

where

$$S(\boldsymbol{Z}) = -\sum_{n=1}^{N} z_n \log z_n \tag{7}$$

is the Shannon entropy and z_n is the n-th component of $\boldsymbol{Z}$. Then, the Interpolation Weight Generator uses $\boldsymbol{L}$ and $D(\boldsymbol{P}, \boldsymbol{U})$ to determine the interpolation weights by

$$\alpha_n = \frac{\exp(D(\boldsymbol{P}, \boldsymbol{U}) \cdot \boldsymbol{L}(\boldsymbol{O}|\boldsymbol{\lambda}_n^{UBM}))}{\sum_{j=1}^{N} \exp(D(\boldsymbol{P}, \boldsymbol{U}) \cdot \boldsymbol{L}(\boldsymbol{O}|\boldsymbol{\lambda}_j^{UBM}))}, \quad \text{for } n = 1, \cdots, N. \tag{8}$$

It is worth noting that, by embedding the divergence measure in Eq. (8), the proposed SD-AKI approach may act as (in two extreme and one normal cases):

(1) a hard-decision *a priori* knowledge-based method, i.e., the weight of the dominant handset will be emphasized and adjusted automatically to approach 1 while all other weights will be de-emphasized and adjusted to approach zero
(2) a blind mismatch-compensation method, i.e., all weights are adjusted to approach $\frac{1}{N}$
(3) a soft-decision interpolation method, i.e., something in between

Lastly, the AKI Interpolator estimates the characteristic of the unknown testing handset by

$$\hat{\boldsymbol{A}}^F = \sum_{n=1}^{N} \alpha_n \boldsymbol{A}_n^F \text{ and } \hat{\boldsymbol{b}}^F = \sum_{n=1}^{N} \alpha_n \boldsymbol{b}_n^F \tag{9}$$

for the feature-space SD-AKI scheme, and by

$$\hat{\boldsymbol{A}}^M = \sum_{n=1}^{N} \alpha_n \boldsymbol{A}_n^M \text{ and } \hat{\boldsymbol{b}}^M = \sum_{n=1}^{N} \alpha_n \boldsymbol{b}_n^M \tag{10}$$

for the model-space SD-AKI scheme.

3. The Compensation of Handset Mismatch

For the feature-space SD-AKI scheme, the handset-mismatch compensation is performed via transforming the input testing feature vector sequence, $\boldsymbol{O}$, by

$$\hat{\boldsymbol{o}}(t) = \hat{\boldsymbol{A}}^F o(t) + \hat{\boldsymbol{b}}^F, \text{ for } t = 1, \cdots, T \tag{11}$$

to obtain a handset mismatch-compensated feature vector sequence, $\hat{\boldsymbol{O}} = \{\hat{\boldsymbol{o}}(1), \cdots, \hat{\boldsymbol{o}}(T)\}$.

For the model-space SD-AKI scheme, we want to transform the set of speaker GMMs, $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_q, q = 1, \cdots, Q\}$ into the testing-speech environment, i.e., $\hat{\boldsymbol{\Lambda}} = \{\hat{\boldsymbol{\lambda}}_q, q = 1, \cdots, Q\}$. Here Q is the total number of speakers in the SID test. In this study, $\boldsymbol{\lambda}_q$ is built from the UBM, $\boldsymbol{\lambda}^{UBM}$, by the MAP adaptation method (Reynolds 2000) using the training speech data, $\boldsymbol{O}_q$, of the q-th speaker of the handset. The handset-mismatch compensation is performed by

$$\hat{\boldsymbol{\mu}}_{q,k} = \hat{\boldsymbol{A}}^M \boldsymbol{\mu}_{q,k} + \hat{\boldsymbol{b}}^M, \; k = 1, \cdots, K \text{ and } q = 1, \cdots, Q, \tag{12}$$

where $\boldsymbol{\mu}_{q,k}$ and $\hat{\boldsymbol{\mu}}_{q,k}$ are the mean vectors of the k-th mixture component of λ_q and its mismatch-compensated counterpart $\hat{\lambda}_q$, respectively. We therefore obtain a set of Q mismatch-compensated speaker GMMs, $\hat{\boldsymbol{\Lambda}} = \{\hat{\lambda}_q = \{(c_k, \hat{\boldsymbol{\mu}}_{q,k}, \boldsymbol{\Sigma}_k), k = 1, \cdots, K\}, q = 1, \cdots, Q\}$.

4. The Handset Mismatch-Compensated Speaker Identification

The handset mismatch-compensated SID is performed last. For the feature-space SD-AKI scheme, the speaker of the input testing speech is identified from the mismatch-compensated feature vector sequence, $\hat{\boldsymbol{O}}$, by the GMM-based SID method using the set of speaker GMMs, $\boldsymbol{\Lambda}$. The optimal speaker identified is determined by

$$\begin{aligned} q^* &= \arg\max_q L(\hat{\boldsymbol{O}} \,|\boldsymbol{\lambda}_q) \\ &= \arg\max_q \sum_{t=1}^{T} \log[P(\hat{\boldsymbol{o}}(t) \,|\boldsymbol{\lambda}_q)\, |J(\hat{\boldsymbol{o}}(t))|], \end{aligned} \tag{13}$$

where $|J(\hat{\boldsymbol{o}}(t))|$ is the determinant of the Jacobian matrix of $\hat{\boldsymbol{o}}(t)$ with respect to $\boldsymbol{o}(t)$ according to the transformation law of probabilities and Eq. (11), (Press, 1994; Sankar and Lee, 1996), i.e.,

$$|J(\hat{\boldsymbol{o}}(t))| = \left|\frac{\partial \hat{\boldsymbol{o}}(t)}{\partial \boldsymbol{o}(t)}\right| = |\hat{A}^F|. \tag{14}$$

For the model-space SD-AKI scheme, the speaker of the input testing speech is identified from the input testing feature vector sequence, $\boldsymbol{O}$, by the GMM-based SID method using the set of handset mismatch-compensated speaker GMMs, $\hat{\boldsymbol{\Lambda}}$. The optimal speaker identified is determined by

$$q^* = \arg\max_q L(\boldsymbol{O} \,|\hat{\boldsymbol{\lambda}}_q) = \arg\max_q \sum_{t=1}^{T} \log P(\boldsymbol{o}(t) \,|\hat{\boldsymbol{\lambda}}_q) \tag{15}$$

5. Discussion

Some advantages of the proposed SD-AKI handset mismatch-compensated SID method can be found. One is that it can be applied to the telephone SID case with very limited available data (a few seconds), because only a small number of parameters have to be estimated. Another is that it can take care of cases of both seen and unseen testing handsets by adjusting interpolation weights automatically without human intervention.

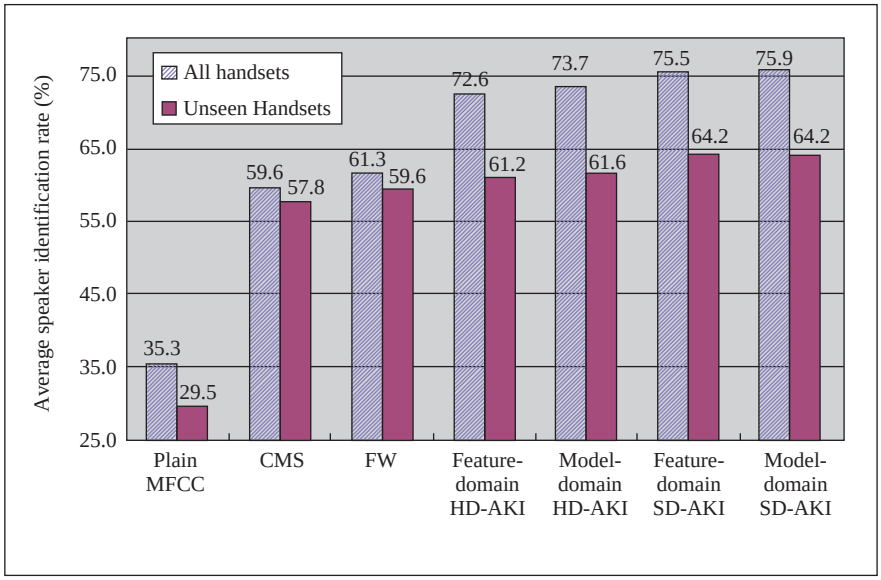
It is also worth noting that all speaker adaptation methods, including SM, FMLLR, MLLR and CMLLR, can not be directly used in the test phase either to transform the testing spectral feature vectors to match with speaker GMMs for the feature-space scheme, or to adapt all speaker GMMs to an unknown testing handset environment for the model-space scheme. The reason is that this will result in the elimination of the characteristic of the testing speaker for the formal case or cause all speaker GMMs to collapse to a UBM model for the latter case, so as to make the following SID fail totally.

III. EXPERIMENTAL RESULTS

1. Experiment Conditions

To evaluate the effectiveness of the proposed SD-AKI handset mismatch-compensated SID method, the handset TIMIT (HTIMIT) database (Reynolds, 1997), which was recorded by the Massachusetts Institute of Technology for the study of the handset mismatch problem, was used. There were in total 384 speakers, each gave ten utterances using a Sennheizer head-mounted microphone (referred to as *sen*). The set of 3840 utterances was played back and recorded through nine other handsets, including four carbon button handsets (referred to as *cb1*, *cb2*, *cb3* and *cb4*), four electret handsets (*el1*, *el2*, *el3* and *el4*), and one portable cordless phone (*pt1*).

All speech signals were first filtered by a bandpass filter with passband 300 ~ 3400 Hz, endpointed



(a)

Scheme		*cb1*	*cb2*	*cb3*	*cb4*	*el1*	*el2*	*el3*	*el4*	*pt1*	Avg.
Plain MFCC		32.0	48.0	4.2	5.3	*52.8*	23.0	55.6	16.6	28.1	29.5
CMS		69.4	71.9	28.1	37.6	*74.7*	62.4	59.6	63.2	53.1	57.8
FE		71.6	72.5	33.2	40.7	*75.6*	63.8	61.3	63.8	54.2	59.6
HD-AKI	Feature-space	75.6	73.9	31.3	50.3	*78.4*	55.6	59.6	66.9	59.6	61.2
	Model-space	75.0	74.2	31.0	49.7	*77.2*	57.9	61.9	67.0	60.1	61.6
SD-AKI	Feature-space	80.1	77.0	31.5	51.1	*83.2*	57.3	66.3	69.7	61.5	64.2
	Model-space	80.1	77.5	31.5	49.4	*82.9*	57.9	66.6	69.4	62.1	64.2

(b)

Fig. 3 A performance comparison between the plain MFCC baseline, CMS or FW, the two proposed SD-AKI schemes and their hard-decision counterparts (HD-AKI): (a) the average SID rates of 9 leave-one-handset-out test turns, (b) the SID rates (%) of unseen handsets in the 9 leave-one-handset-out test turns. Here, *cb1~4*, *el1~4*, and *pt1* shown in the 1st row of (b) are the unseen handsets in these 9 leave-one-handset-out test turns

using energy-based rules and then processed to extract 38-dimensional spectral feature vectors, composed of 12 mel-frequency cepstral coefficients (MFCCs), 12 Δ-MFCCs, 12 Δ^2-MFCCs, Δ-log energy, and Δ^2-log energy. A frame size of 30 ms with 10-ms shift was used. Moreover, low energy frames were dropped.

In all our experiments, a subset of HTIMIT comprising 356 speakers (178 females and 178 males) was used. Each speaker had ten 16-second and another (disjoint) ten 4-second segments from all ten different handsets for testing, respectively.

In all cases, the 16-second speech segments of all speakers from *sen* were taken as the data. They were first collectively used to train a 256-mixture UBM, $\boldsymbol{\lambda}^{UBM}$, and then separately used to adjust the UBM by the MAP adaptation method (Reynolds *et al.*, 2000) to generate the set of 356 speaker GMMs, $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_q, q = 1, \cdots, 356\}$.

To simulate the scenario of variable unseen testing handsets to evaluate the proposed SD-AKI schemes, a leave-one-handset-out (excepted *sen*) experimental strategy was adopted. There were in total 9 leave-one-handset-out turns. For each turn, we chose 9 seen handsets (including *sen*) and one unseen handset in turn. The experiment setting is described in more detail as follows.

In the training phase, the 16-second speech segments of all speakers from the nine seen handsets were used to construct the *a priori* knowledge of handset characteristics as well as the set of handset-specific UBMs. In the test phase, we took other 4-second (disjoint with the training material) speech segments of all speakers from all ten handsets comprising 9 seen handsets (including *sen*) and one unseen handset, as testing utterances. So, in each leave-one-handset-out turn we had 356 and 356*9 speaker identification trials from one unseen and 9 seen handset, respectively.

Before we evaluated the proposed SD-AKI method, the plain MFCC, CMS, FW methods and its hard-decision counterpart (HD-AKI) were first tested and taken as the baselines.

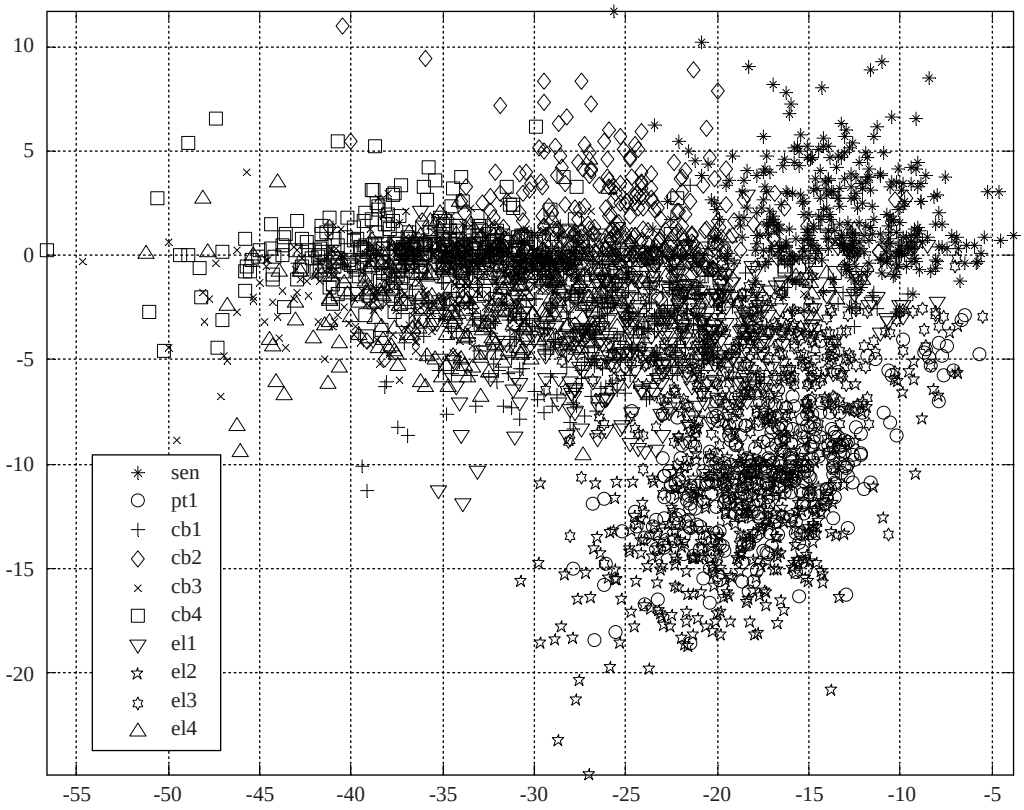


Fig. 4 The scattering plot of the 1st vs. 2nd dimensions of observed average plain MFCCs of all speakers when their speeches pass through ten different handsets

2. Plain MFCC Baseline and Handset Mismatch Analysis

First of all, if training and test were all executed on the same *sen* handset (handset match case or seen handset), the speaker identification rate was found to be 87.5%. However, if training and test handsets may be different (handset mismatch or unseen handset condition), i.e., training on *sen* but test on *sen*, *cb1~4*, *el1~4* and *pt1*, the average speaker identification rates was dramatically degraded to 35.3%. The result further decreased to 29.5% if we only counted the results of the 9 unseen handsets. The detail experimental results of the plain MFCC features baseline could be found in Figs. 3(a) and 3(b).

The reason why handset mismatch may cause so serious performance degradation on HTIMIT corpus was analyzed. Fig. 4 displays the scattering plot of the first two dimensions of the observed average plain MFCCs of each speaker when speech signals of all speakers passed through the ten handsets used. It can be seen from the figure that the distribution of the observed feature vectors of all speakers could be greatly changed when they used different handsets. This change in distribution, of course, will cause a mismatch between the testing utterance and the speaker models trained from *sen* and hence might result in serious performance degradation.

3. CMS and FW

CMS and FW were then utilized to partially alleviate the handset mismatch problem. In both the training and test phase, the cepstral mean vector or the distribution of MFCCs of each input speech segment was removed or normalized into a Gaussian distribution first using CMS or FW, respectively.

The experimental results are also displayed in Figs. 3(a) and 3(b). It can be seen from the figure that the average speaker identification rates for all 10 handsets was raised to 59.6% and 61.3% for CMS and FW, respectively, as we counted the results of both seen and unseen handsets. If we only counted the results of the 9 unseen handsets, the results were also increased to 57.8% and 59.6% for CMS and FW, respectively. These results reveal the benefits of CMS and FW for SID in handset mismatch condition.

4. The Proposed SD-AKI Method and its Hard-Decision Counterpart

We then examined the performances of the proposed feature- and model-space SD-AKI schemes and their hard-decision counterparts (HD-AKI). All

experimental settings and handset compensation operations were same for both SD-AKI and HD-AKI except that HD-AKI only picked up and utilized the characteristic of the most possible handset. It is also worth noting that the average accuracy for HD-AKI to choose the correct seen handset is 95.3%.

The experimental results are displayed in Figs. 3(a) and 3(b). First of all, it can be seen from the figures that feature-/model-space HD- and SD-AKI schemes dramatically raised the average identification rates to 72.3%/73.7% and 75.5%/75.9%, respectively, if we counted the results of both seen and unseen handsets. It can also be found from Figs. 3(a) and 3(b) that the average identification rates were also increased to 61.2%/61.6% and 64.2%/64.2%, respectively, for feature-/model-space HD- and SD-AKI schemes if we only counted the results of unseen handsets.

These results revealed that (1) the benefits of *a priori* handset knowledge-based approaches (HD- and SD-AKIs), (2) model-space approaches were slightly better than the feature-space ones in both HD- and SD-AKI cases and (3) hard-decision-based approaches were more sensitive to the disturbance of unseen handsets (may due to handset detection error).

Comparing with the result of the CMS, FW and HD-AKI methods, both the two proposed SD-AKI schemes performed much better. We can therefore conclude that the feature- and model-space SD-AKI schemes are more effective on compensating the handset mismatch so as to improve the SID performances for both cases of all handsets and unseen handsets.

5. Discussions

Three important issues on the proposed SD-AKI approach including (1) the benefit of introducing the divergence measure, (2) handset coverage and (3) computation cost are further addressed in this subsection.

Since the number of available seen handsets is usually limited, the set of pre-collected *a priori* handset knowledge may be a sparse representation of handset space. If only the normalized likelihoods (*a posteriori* in Eq. (5)) are used to generate the interpolation weights, the pre-collected handset characteristics may be smear out too much, especially, for those test handsets similar or same to the seen ones. According to some preliminary experiments, the divergence measure did bring significant benefit for those seen test handsets.

Although, the proposed SD-AKI method in general performed better than CMS and FW approaches in the leave-one-handset-out experiments, especially, for those seen handsets. However, it was found from Fig. 3(b) that its performance on unseen handset *el2* was less than CMS and FW methods. This may due to the incomplete coverage of the set of pre-prepared *a priori* handset knowledge, i.e., other seen handsets may not close enough to handset *el2* (see Fig. 4). This problem may be partially alleviated by further recruiting the characteristics of CMS- or FW-based features or models into the set of the *a priori* handset knowledge.

Comparing with other approaches, the proposed SD-AKI requires extra computation power on calculating the handset *a posteriori* probabilities (Eq. (4)). However, the most computation intensive procedure of SID is usually on the evaluating the likelihoods of a large amount of speaker models (Eqs. (13) or (15)). Since the number of handsets is often one to two order less than the number of speaker models, this overhead may be acceptable.

IV. CONCLUSIONS

A new SD-AKI method for handset mismatch-compensated speaker identification was proposed in this paper. Two schemes of realizing the SD-AKI method in feature and model spaces were discussed. Experimental results on the HTIMIT database have showed that both feature- and model-space SD-AKI schemes performed much better than the CMS, FW and their hard-decision counterpart methods on both cases of all-handset and unseen-handset SIDs. So, the proposed SD-AKI-based method is promising for robust speaker identification over PTSN.

ACKNOWLEDGEMENT

This work was supported by the National Science Council, Taiwan, under contract NSC 96-2221-E-027-100-MY2.

NOMENCLATURE

$\boldsymbol{A}_n^F$	feature transformation matrix for the n-th seen handset
$\hat{\boldsymbol{A}}^F$	estimated feature transformation matrix
$\boldsymbol{A}_n^M$	model transformation matrix for the n-th seen handset
$\hat{\boldsymbol{A}}^M$	estimated model transformation matrix
$\boldsymbol{b}_n^F$	feature transformation bias vector for the n-th seen handset
$\hat{\boldsymbol{b}}^F$	estimated of feature transformation bias vector
$\boldsymbol{b}_n^M$	model transformation bias vector for the n-th seen handset
$\hat{\boldsymbol{b}}^M$	estimated of model transformation bias vector
c_k	mixture weight of the k-th mixture component of GMMs
$D(\boldsymbol{P}, \boldsymbol{U})$	Jensen-Shannon divergence measure

	between two distributions $\boldsymbol{P}$ and $\boldsymbol{U}$
$\boldsymbol{H}$	*a priori* handset knowledge
$\boldsymbol{H}^F$	feature-space *a priori* handset knowledge
$\boldsymbol{H}^M$	model-space *a priori* handset knowledge
$\boldsymbol{h_n}$	characteristics of the n-th seen handset
$\hat{\boldsymbol{h}}$	estimated handset characteristics
$\boldsymbol{J}(\hat{\boldsymbol{o}}(t))$	Jacobian matrix of $\hat{\boldsymbol{o}}(t)$ with respect to $\boldsymbol{o}(t)$
K	number of mixture components
$\boldsymbol{L}$	log-likelihood vector
$L(\boldsymbol{O}\|\lambda_n^{UBM})$	log-likelihood function given n-th seen handset UBM
N	number of seen handsets
n	index of seen handsets
$\boldsymbol{O}$	speech feature vectors
$\boldsymbol{O}_n$	speech feature vectors from the n-th seen handset
$\hat{\boldsymbol{O}}$	compensated speech feature vectors
$o(t)$	speech feature vector of the t-th frame
$o_n(t)$	speech feature vector of the t-th frame of the n-th seen handset
$\hat{\boldsymbol{o}}(t)$	compensated speech feature vectors of the t-th frame
$\boldsymbol{P}$	*a posteriori* probability vector
$p(\boldsymbol{\lambda}_n^{UBM}\|\boldsymbol{O})$	*a posteriori* probability of n-th handset UBM given the speech feature vectors
$p(\boldsymbol{o}(t)\|\boldsymbol{\lambda}_n^{UBM})$	probability of speech feature vector of the t-th frame given n-th seen handset UBM
Q	number of speakers
q^*	identified speaker
$S(\boldsymbol{Z})$	Shannon entropy of distribution $\boldsymbol{Z}$
T	number of frames of an input utterance
t	index of frame
$\boldsymbol{Z}$	probability distribution
z_n	n-th component of $\boldsymbol{Z}$
$\boldsymbol{U}$	uniform probability mass function
$\boldsymbol{\alpha}$	interpolation weight vectors
α_n	n-th interpolation weight
$\boldsymbol{\Lambda}$	set of speaker models
$\hat{\boldsymbol{\Lambda}}$	set of compensated speaker models
$\boldsymbol{\Lambda}^{UBM}$	set of universal background models
$\boldsymbol{\lambda}_q$	q-th speaker model
$\hat{\lambda}_q$	compensated speaker model
λ^{UBM}	universal background model of the training handset
λ_n^{UBM}	universal background model of the n-th seen handset
$\boldsymbol{\Sigma}_k$	covariance matrix of the k-th mixture component
$\boldsymbol{\mu}_k$	mean vector of the k-th mixture component of UBM
$\boldsymbol{\mu}_{n,k}$	mean vector of the k-th mixture component of n-th seen handset UBM
$\boldsymbol{\mu}_{q,k}$	mean vectors of the k-th mixture component of the q-th speaker model
$\hat{\boldsymbol{\mu}}_{q,k}$	compensated mean vectors of the k-th mixture component of the q-th speaker model

REFERENCES

Adami, A. G., Mihaescu, R., Reynolds, D. A., Godfrey, J. J., 2003, "Modeling Prosodic Dynamics for Speaker Recognition," *Proceeding of 2003 IEEE International Conference on Acoustics, Speech, And Signal Processing*, Hong Kong, pp. 788-791.

Bates, R. A., and Ostendorf, M., 1999, "Reducing the Effects of Linear Channel Distortion on Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 1, Issue 5, pp. 594-597

Beaufays, F., and Weintraub, M., 1997, "Model Transformation for Robust Speaker Recognition from Telephone Data," *Proceeding 1997 of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, pp. 1063-1066.

Burbea, J., and Rao, C. R., 1982, "On the Convexity of Some Divergence Measures Based on Entropy Functions," *IEEE Transactions on Information Theory*, Vol. 28, No. 3, pp. 489-495.

Campbell, J. P., Jr., 1997, "Speaker Recognition: a Tutorial," *Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1437-1462.

Chaudhari, U. V., Navratil, J., Maes, S. H., 2003, "Multigrained Modeling with Pattern Specific Maximum Likelihood Transformations for Text-Independent Speaker Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 1, pp. 61-69.

Chen, Z. H., Liao, Y. F., and Juang, Y. T., 2005, "Prosody Modeling and Eigen-Prosody Analysis for Robust Speaker Recognition," *Proceeding of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, pp. 185-188.

Chien, J. T., Wang, H. C., 1998, "Phone-dependent Channel Compensated Hidden Markov Model for Telephone Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 5, Issue 6, pp. 143-145

Digalakis, V. V., Rtischev, D., and Neumeyer, L. G., 1995, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 5, pp. 357-366.

Faundez-Zanuy, M., Monte-Moreno, E., 2005, "State-of-the-art in Speaker Recognition," *IEEE Aerospace and Electronic Systems Magazine*, Vol. 20, No. 5, pp. 7-12.

Furui, S., 1981, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Transactions on Speech and Audio Processing*, Vol. 29, No. 2, pp. 254-272.

Gales, M. J. F., and Woodland, P. C., 1996, "Mean and Variance Adaptation within the MLLR Framework," *Computer Speech and Language*, Vol. 10, No. 4, pp. 249-264.

Gales, M. J. F., 1997, "Maximum Likelihood Linear Transformations for HMM-Based Speech Recognition," *Technical Report*, Cambridge University Engineering Department, UK.

Gong, Y., 2005, "A Method of Joint Compensation of Additive and Convolutive Distortions for Speaker-Independent Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, Issue 5, Part 2, pp. 975-983

Hermansky, H., and Morgan, N., 1994, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 4, pp. 578-589.

Hung, J. W., Lee, L. S., 2006, "Optimization of Temporal Filters for Constructing Robust Features in Speech Recognition," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 3.

Jiang, H., and Deng, L., 2000, "A Robust Compensation Strategy for Extraneous Acoustic Variations in Spontaneous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 10, Issue 1, pp. 9-17

Juang, B. H., and Rahim, M. G., 1996, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 4, Issue 1, pp. 19.

Kozat, S. S., Visweswariah, K., and Gopinath, R., 2007, "Efficient, Low Latency Adaptation for Speech Recognition," *Proceeding of 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, Hawaii, U.S. A., pp.777-780.

Mak, M. W., Kung, S. Y., 2002, "Combining Stochastic Feature Transformation and Handset Identification for Telephone-Based Speaker Verification," *Proceeding of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, pp.701-704.

Mammone, R., Zhang, X., and Ramachandran, R., 1996, "Robust speaker recognition – a feature-based method," *IEEE Signal Processing Magazine*, Vol. 1, No. 2, pp. 58-71.

Mason, M., Vogt, R., Baker, B., and Sridharan, S., 2005, "Data-Driven Clustering for Blind Feature Mapping in Speaker Verification," *Proceeding of the 9th European Conference on Speech Communication and Technology*, Lisboa, Portugal, pp. 3109-3112.

Murthy, H. A., Beaufays, F., Heck, L. P., and Weintraub, M., 1999, "Robust Text-Independent Speaker Identification over Telephone Channels," *IEEE Transactions on Speech Audio Processing*, Vol.7, No. 5, pp. 554-568.

Pelecanos, J., and Sridharan, S., 2001, "Feature Warping for Robust Speaker Verification," *Proceeding of 2001: A Speaker Odyssey – The Speaker Recognition Workshop*, Crete, Greece, Paper #1038.

Quatieri, T. F., Reynolds, D. A., and O'Leary, G. C., 2000, "Estimation of Handset Nonlinearity with Application to Speaker Recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 5, pp. 567-584.

Reynolds, D. A., Rose, R. C., 1995, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, Vol. 3, No. 1, pp. 72-83.

Reynolds, D. A.,: 1997, "HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects," *Proceeding of IEEE 1997 International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, pp. 1535-1538.

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B., 2000, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, Vol. 10, No. 1-3, pp. 19-41.

Reynolds, D. A., 2003, "Channel Robust Speaker Verification via Feature Mapping," *Proceeding of 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, pp. 53-56.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., 1994, "Numerical Recipes in C, The Art of Scientific Computing," Second Edition, *Cambridge Press*, New York, USA, pp. 287-288.

Sankar, A., and Lee, C. H., 1996, "A Maximum-Likelihood Method to Stochastic Matching for Robust Speech Identification," *IEEE Transactions on Speech and Audio Processing*, Vol. 4, No. 3, pp. 190-202.

Sant'Ana, R., Coelho, R., and Alcaim, A., 2006, "Text-Independent Speaker Recognition Based on the Hurst Parameter and the Multidimensional Fractional Brownian Motion Model," *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 3, pp. 931-940.

Schütze, H., and Manning, C. D., 1999, *Foundations*

of Statistical Natural Language Processing, The MIT Press, Cumberland, Rhode Island, USA, pp. 304.

Teunen, R., Shahshahani, B., and Heck, L. P., 2000, "A Model based Transformational Method to Robust Speaker Identification," *Proceeding of the Sixth International Conference on Spoken Language Processing*, Beijing China, pp. 495-498.

Xiang, B., Chaudhari, U. V., and Navratil, J., 2002, "Ramaswamy, R. A. Gopinath, "Short-Time Gaussianization for Robust Speaker Verification," *Proceeding of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, pp. 681-684.

Yang, J. H., and Liao, Y. F., 2004, "Unseen Handset Mismatch Compensation Based on Feature/Model-Space *A priori* Knowledge Interpolation for Robust Speaker Identification," *Proceeding of the 4th International Symposium on Chinese Spoken Language Processing*, Hong Kong, pp. 65-68.

Yiu, K. K., Mak, M. W., Cheung, M. C., Kung, and S. Y., 2004, "A New Method to Channel Robust Speaker Verification via Constrained Stochastic Feature Transformation," *Proceeding of the Eighth International Conference on Spoken Language Processing*, Jeju Island, Korea, pp. 1753-1756.

Yuo, K. W., Hwang, T. H., and Wang, H. C., 2005, "Combination of Autocorrelation-based Features and Projection Measure Technique for Speaker Identification," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 4, pp. 565-574.

Zhao, Y., 2000, "Frequency-domain Maximum Likelihood Estimation for Automatic Speech Recognition in Additive and Convolutive Noises," *IEEE Transactions on Speech and Audio Processing*, Vol. Issue 3, pp. 255-266.

Manuscript Received: Feb. 21, 2008
Revision Received: Sep. 19, 2008
and Accepted: Oct. 19, 2008