APPLYING ARTICULATORY FEATURES TO TELEPHONE-BASED SPEAKER VERIFICATION

Ka-Yee Leung, 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

This paper presents an approach that uses articulatory features (AFs) derived from spectral features for telephone-based speaker verification. To minimize the acoustic mismatch caused by different handsets, handset-specific normalization is applied to the spectral features before the AFs are extracted. Experimental results based on 150 speakers using 10 different handsets show that AFs contain useful speaker-specific information for speaker verification and the use of handset-specific normalization significantly lowers the error rates under the handset mismatched conditions. Results also demonstrate that fusing the scores obtained from an AF-based system with those obtained from a spectral feature-based (MFCC) system helps lower the error rates of the individual systems.

1. INTRODUCTION

Most traditional speaker recognition systems are based on the modeling of short-term spectral information [1]. The advantage of using short-term spectral information is that promising results are obtainable from a limited amount of training data. In recent years, researchers have started to investigate the high-level speaker information, such as the usage or duration of particular words, prosodic features, etc., that is obtainable from speech [2]. These high-level features are generally neglected by the traditional spectral feature-based systems. However, recent research has shown that when high-level features are combined with low-level features (spectral features), significant improvement in speaker recognition accuracy can be obtained [2].

Speech is produced by the continuous movements of articulators in the vocal tract excited by the air stream originated from the lung. These speaker-characterized articulations and excitations, which imparted to the produced speech, are the origin of unique speaker information [3]. However, articulatory information has not been widely applied to automatic speaker recognition because the extraction of these features is not trivial.

In this paper, we explore the use of articulatory features (AFs) to capture the movements of articulators in the vocal tract and their excitation during sound production for speaker verification. AFs are the abstract representations of some important speech production properties, such as the manner and place of articulation, the vocal cord excitation, and lip motion, etc. AFs have been adopted as an alternative or supplementary features for speech recognition [4, 5], language ID [6] and confidence measure [7]. Preliminary work of applying AF on speaker ID have been performed by K. Kirchhoff [8] in the 2002 JHU Summer Workshop on Human Language Technology. It was found that AFs are complementary to spectral features and that better performance can be obtained when they are used together.

This work was supported by The Hong Kong Polytechnic University, Grant No. A-PE44 and Research Grant Council of the Hong Kong ASR (Project No. CUHK 1/02C).

Articulatory properties	*Classes*	*# Class*
Voicing	Voiced, Unvoiced	2
Front-back	Front, Back ,Nil	3
Rounding	Rounded, Not Rounded, Nil	3
Manner	Vowel, Stop, Fricative, Nasal Approximant-Lateral	5
Place	High, Middle, Low, Labial, Dental Coronal, Palatal, Velar, Glottal	9

Table 1. Five articulatory properties and the number of classes in each properties.

2. ARTICULATORY FEATURE EXTRACTION

To extract AFs from speech, a set of articulatory classifiers are trained to learn the mapping between the acoustic signals and the articulatory states. Either images, such as X-ray that records the actual articulatory positions, or mappings between phonemes and their corresponding articulatory properties can be used to train the classifiers. In this work, AFs were extracted from acoustic signals based on an approach similar to [5]. Specifically, to obtain the AFs, a sequence of acoustic vectors are fed to five classifiers in parallel, where each classifier represents a different articulatory property. The outputs of these classifiers (the posterior probabilities) are concatenated to form the AF vectors. The extracted AFs can be treated as an intermediate representation of the acoustic signals.

In our verification system, five different articulatory properties, as tabulated in the first column of Table 1, are used. For each property, a Multi-Layer Perceptron (MLP) is used to estimate the probability distributions of its predefined output classes (they are listed in the second column of Table 1). The extraction process is illustrated in the right dotted box of Fig. 1.

The inputs to these five AF-MLPs are identical while their numbers of outputs are equal to the numbers of AF classes listed in the last column of Table 1. To ensure a more accurate estimation of the AF values, multiple frames of Mel-frequency cep-

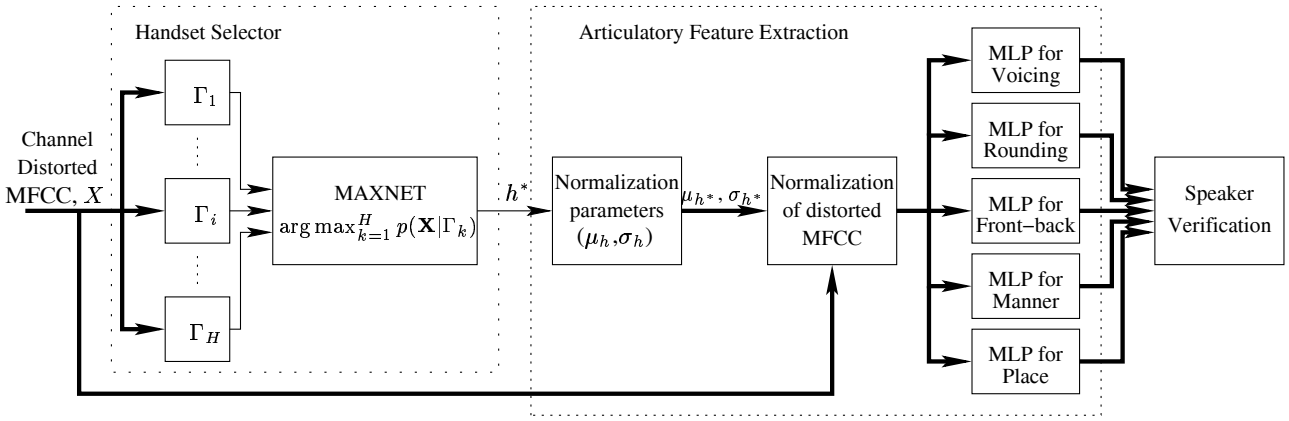

Fig. 1. Combination of handset identification and AF extraction for robust speaker verification. Arrows with thick lines are inputs/outputs with multiple dimensions. Each AF-MLP takes 9 consecutive, 12-dimensional normalized MFCCs as input. The number of outputs for the Voicing, Rounding, Front-back, Manner and Place MLPs are 2, 3, 3, 5 and 9 respectively.

stral coefficients (MFCCs), which are $[t - \frac{n}{2}, \ldots, t, \ldots, t + \frac{n}{2}]$ of MFCCs created by a moving window, are served as the inputs to the AF-MLPs at frame t. Rather than feeding the MFCCs directly to the AF-MLPs, they are normalized to zero mean and unit variance. The normalization parameters, a mean vector μ and a standard deviation vector σ each with dimension the same as that of the MFCCs, are obtained globally. Given an MFCC vector $\mathbf{x}_t$, the normalization for dimension i is done by applying

$$x_t^{norm}(i) = \frac{x_t(i) - \mu(i)}{\sigma(i)}. \quad (1)$$

The normalization aims to remove the variations of input features among different dimensions so that the determination of MLP weights is not dominated by those input features with large magnitude.

The AF-MLPs can be trained from speech data with time-aligned phonetic labels. The alignments can be obtained from transcriptions or from the Viterbi decoding using phoneme models. With the phoneme labels, articulatory classes can then be derived from a mapping between phonemes and their states of articulations [5].

3. ROBUST SPEAKER VERIFICATION

The procedure of AF-based speaker verification is illustrated in Fig. 1. It can be divided into three steps: handset identification, AF extraction and speaker verification.

3.1. Handset Identification

Previous work on telephony speech has shown that using different types of handsets can cause various degrees of distortion to the speech signals [11]. Therefore, the AFs extracted from the distorted features will become less reliable and cannot match well with the speaker and background models that were trained using data recorded from a different handset.

To reduce the mismatch between training and testing conditions, the handset of each testing utterance is identified and handset-specific compensation is applied to the distorted MFCCs. We have adopted our recently proposed handset selector [9, 10] to identify the most likely handset given a testing utterance. The handset selector is shown in the left dotted box of Fig. 1. Before verification, H handset specific Gaussian Mixture Models (GMMs), $\{\Gamma_k\}_{k=1}^{H}$, were obtained offline, where each Γ_k was trained using the distorted speech collected from the telephone handset k. During verification, the most likely handset h^* of every testing utterance is identified by feeding the distorted MFCCs, $\mathbf{X}= \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ to $\{\Gamma_k\}_{k=1}^{H}$.

3.2. MFCC Normalization and AF Extraction

The AF values determined from the AF-MLPs are closely related to the quality of MFCCs as they are the source of the AF extraction. The AFs cannot be correctly estimated if the AF-MLPs take the distorted MFCCs as their inputs. To compensate for the distortion caused by different handsets, normalization parameters (μ_h, σ_h) for each handset h are determined. During verification, the normalization parameters corresponding to the identified handset are used to normalize the distorted MFCCs according to (1). The normalized MFCCs are then fed to the five AF-MLPs to determine the AFs. So, the variation of MFCCs due to handsets difference can be minimized and they are transformed to a range which is closer to the training patterns.

3.3. Speaker Verification

According to Table 1, there are a total of 22 articulatory classes, which result in a 22-dimensional AF vector for each frame. For each testing utterance from a claimant, a sequence of 22-dimensional AF vectors $\mathbf{Y}$were fed to a speaker model ($\mathcal{M}_s$) and a background model ($\mathcal{M}_b$) to obtain a verification score $S(\mathbf{Y})$

$$S(\mathbf{Y}) = \log p(\mathbf{Y}|\mathcal{M}_s) - \log p(\mathbf{Y}|\mathcal{M}_b). \quad (2)$$

$S(\mathbf{Y})$ was compared with a threshold to make a decision. The threshold was varied to obtain a speaker-independent equal error rate (EER), i.e., the point at which the false rejection rate is equal to the false acceptance rate.

Speaker Set (50 female and 50 male)	Impostor Set (25 female and 25 male)
fadg0,faem0,. . .,fdxw0,mabw0, majc0,. . .,mfgk0,mjls0,mjma0, mjmd0,mjmm0,mpdf0	feac0,fear0,. . .,fjem0 mfxv0,mgaw0,. . .,mjlg1

Table 2. Speaker identities in the speaker set and the impostor set. The speakers in these sets were arranged alphabetically.

4. EXPERIMENTS AND RESULTS

4.1. Corpus

In this work, the HTIMIT corpus [12] was used for performance evaluation. HTIMIT was constructed by playing a gender-balanced subset of the TIMIT corpus through 9 telephone handsets and a Sennheizer head-mounted microphone. This set-up introduces real handset-transducer distortion in a controlled manner but without losing the time-aligned phonetic transcriptions of the TIMIT corpus. This feature makes HTIMIT ideal for studying the handset variability in speech and speaker recognition systems [13]. It also facilitates the training of AF-MLPs by mapping the time-aligned phoneme labels to their corresponding articulatory classes.

4.2. Speaker Enrollment

Two disjointed gender-balanced speaker set and imposter set, which consists of 100 and 50 speakers respectively, were selected from the HTIMIT corpus. The speaker identities are listed in Table 2.

For the system that uses spectral features only (referred to as MFCC system hereafter), 12-th order MFCCs were computed every 14ms using a Hamming window of 28ms. For the system that uses AFs as features (referred to as AF system hereafter), 22-dimensional AF vectors were obtained from the five AF-MLPs, each with 108 input nodes (9 frames of 12-dimensional MFCCs) and 50 hidden nodes. The MLPs were trained using the Quicknet [14]. Training data includes all sentences collected using the head-mounted microphone (senh) of all speakers in HTIMIT, excluding speakers from the speaker and imposter sets.

For each system, a 64-center universal background model $\mathcal{M}_b$ was trained using the SA and SX sentences from all speakers in the speaker set. For each speaker s in the speaker set, a speaker model $\mathcal{M}_s$ was adapted from $\mathcal{M}_b$ using MAP adaptation [1]. The SA and SX sentences collected using the head-mounted microphone (senh) were used for enrollment and adaptation. $\{\mathcal{M}_s, \mathcal{M}_b\}$ were tested on the SI sentences of speaker s and the 50 impostors.

4.3. Robustness Enhancement

For the MFCC system, either cepstral mean subtraction (CMS) or stochastic feature transformation (SFT) [9] was adopted to enhance the robustness of handset mismatch during verification. When applying SFT during verification, the handset selector, $\{\Gamma_k\}_{k=1}^{H}$, was first applied to the testing utterance to identify the handset. The SFT parameters of the identified handset were then used to transform the distorted MFCC vectors. The SFT parameters for each handset appeared in the handset selector were estimated from the SX and SA sentences of 10 speakers (first ten speakers in the speaker set) using the corresponding handset.

The handset selector in the AF system is also based on MFCCs, and $\{\Gamma_k\}_{k=1}^{H}$ are identical to those of the MFCC system. Both the verification performance of AFs extracted from MFCCs and from MFCCs with CMS were evaluated. As MFCCs need to be normalized before feeding to the AF-MLPs, we also evaluate the AF system when CMS and HS-Norm are used together. If CMS is applied before HS-Norm, accuracy of the handset selector may be lowered as the channel characteristics in the MFCCs will be removed. Due to this, HS-Norm is applied before CMS. For each type of handsets, the normalization parameters (μ, σ) were estimated from the same data used to estimate the SFT parameters.

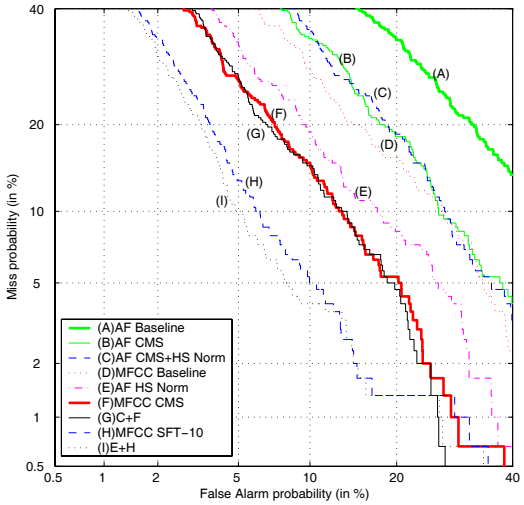

Fig. 2. DET plots of different features and channel compensation approaches based on the testing utterances from handset el3. For ease of comparison, methods in the legend are arranged in descending order of EER.

4.4. System Fusion

Although both the AF and MFCC systems take MFCCs as input, they attempt to capture two different information from the speech signals. The MFCC system attempts to capture the acoustic characteristics while the AF system attempts to capture the articulatory properties. Therefore, fusion of these two systems should obtain performance better than the individual systems.

In this work, utterance scores, as given in (2), obtained from the MFCC system and the AF system were linearly combined to produce the fusion scores

$$S_F(\mathbf{Y}_{mfcc}, \mathbf{Y}_{af}) = (1 - w_{af})S(\mathbf{Y}_{mfcc}) + w_{af}S(\mathbf{Y}_{af}), \quad (3)$$

where w_{af} is a handset-specific fusion weight. It was determined from data used for estimating the normalization parameters and the SFT parameters of each handset type.

4.5. Results

Table 3 summaries the EERs obtained from the approaches discussed above. The verification results of MFCC system, AF system and the fusion of the two systems are listed in Rows 1-3, Rows 4-7 and Rows 8-9, respectively. The EERs are the average of all SI sentences from 100 speakers and 50 impostors. The DET plots corresponding to handset el3 are shown in Fig. 2.

Evidently, for the MFCC system, *MFCC SFT* outperforms *MFCC Baseline* and *MFCC CMS*. For the AF system, verifications

Row	Features	Equal Error Rate (%)										
		cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average	senh
1	MFCC Baseline	16.35	23.68	30.20	28.86	14.48	27.64	16.84	23.90	32.61	**23.84**	2.93
2	MFCC CMS	8.31	7.48	28.65	21.04	8.29	9.76	11.65	8.86	11.13	**12.80**	5.47
3	MFCC SFT	5.30	4.33	19.90	14.84	4.68	6.63	7.55	4.10	7.45	**8.31**	2.95
4	AF Baseline	23.36	30.80	38.90	34.25	24.40	34.32	26.20	27.30	38.87	**30.93**	6.60
5	AF HS-Norm.	10.11	9.97	26.28	20.48	11.69	13.95	13.22	8.01	17.42	**14.57**	6.56
6	AF CMS	15.84	14.67	35.18	30.01	15.16	17.82	19.16	15.18	23.51	**20.73**	11.22
7	AF CMS + HS-Norm.	14.50	14.45	28.65	23.65	14.97	17.66	18.98	13.51	19.17	**18.39**	11.22
8	3 + 5	4.87	3.91	19.24	13.92	4.68	6.30	6.59	3.67	7.49	**7.85**	2.85
9	2 + 7	7.64	7.54	25.55	18.57	8.20	9.21	11.33	8.19	9.64	**11.76**	5.35

Table 3. Equal error rates based on different approaches and different handsets. *3+5* is the fusion of *MFCC SFT* and *AF HS-Norm* while *2+7* is the fusion of *MFCC CMS* and *AF CMS + HS-Norm*. CMS, SFT and HS-Norm stands for cepstral mean subtraction, stochastic feature transformation [9], and handset normalization, respectively. Fusions based on other combinations have also been performed; however, we only list the two which give the best results. The handset recognition accuracy is 98.35%. Note that the *MFCC baseline*, *MFCC CMS*, *AF Baseline* and *AF CMS* do not require the handset selector.

based on the AFs extracted from the MFCCs and from the MFCCs with CMS are respectively named as *AF Baseline* and *AF CMS* in Table 3. Applying handset normalization (HS-Norm in Table 3) significantly reduces the average EER from 30.93% to 14.57% for *AF Baseline* and 20.73% to 18.39% for *AF CMS*. This represents an error reduction of 52.89% and 11.29% for *AF Baseline* and *AF CMS* respectively. Note that the results of SFT or HS-Norm give the upper bound performance of their handset compensation ability. As the HTIMIT corpus adopts a parallel recording approach and there is a single recording session for each handset type, the transformations obtained from SFT or HS-Norm can match well with the handset distortion. It is of interest to investigate the situation in which the variations in the telephone line distortion are also considered.

The aim of applying HS-Norm is to shift the channel distorted MFCCs back to a range comparable to the normalized MFCCs used for MLP training, so that the five AF-MLPs can work well on both channel matched and channel mismatched MFCCs. This objective was largely achieved because the EERs of different handsets under *AF HS-Norm* were made closer to the EER obtained from *AF Baseline* using the enrollment handset (senh). As CMS removes most of the channel characteristics, *AF CMS+HS-Norm* achieves a more significant EER reduction on several handsets only, e.g., cb3, cb4, el4 and pt1. As a result, the average error reduction becomes less significant.

The individual MFCC and AF systems that give the lowest EER were fused together. When no CMS was used, fusing *MFCC SFT* and *AF HS-Norm* reduces the average EER from 8.31% to 7.85%, which represents a 5.54% error reduction. When CMS was used, the fusion of *MFCC CMS* and *AF CMS+HS-Norm* lowers the average EER from 12.80% to 11.76%, which represents a 8.13% error reduction. This suggests that the acoustic characteristics represented by the MFCCs and the articulatory properties represented by the AFs are partially complementary, although they are from the same source.

5. CONCLUSIONS

This paper has presented a speaker verification approach using the articulatory features (AFs) derived from MFCCs. Results based on 150 speakers from HTIMIT have shown that AFs contain speaker-specific information which is useful for speaker verification. In order to increase the robustness of AFs to channel mismatch, handset-specific normalization was applied during AF extraction. Results show that the normalization significantly reduces the equal error rate under the handset mismatch conditions. In addition, when the proposed AF system is fused with a traditional MFCC system, an equal error rate lower than that of the individual systems is obtained.

6. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, pp. 19- 41, 2000.

[2] D. A. Reynolds et. al., "The superSID project: exploiting high-level information for high-accuracy speaker recognition," *Proc. ICASSP'03*, vol. IV, pp.784-787, 2003.

[3] G. R. Doddington. "Speaker Recognition-Identifying People by their Voices," *Proc. IEEE*, vol. 73, pp. 1651-1664, 1985.

[4] K. Erler and L. Deng, "Hidden Markov Model Representation of Quantized Articulatory Features for Speech Recognition". *Computer Speech and Language*, Vol. 7, no 3, pages 265-282, 1993.

[5] K. Kirchhoff, "Robust Speech Recognition Using Articulatory Information," PhD thesis, University of Bielefeld, 1999.

[6] S. Parandekar and K. Kirchhoff, "Multi-Stream Language Identification Using Data-driven Dependency Selection," *Proc. ICASSP'03*, Vol. I, pages 28-31, 2003.

[7] K. Y. Leung and M. Siu, "Phone level confidence measure using articulatory features," *Proc. ICASSP'03*, Vol. I, pages 600-603, 2003.

[8] JHU WS'2002 SuperSID group website
http://www.clsp.jhu.edu/ws2002/groups/supersid/

[9] M. W. Mak and S. Y. Kung, "Combining stochastic feature transformation and handset identification for telephone-based speaker verification," *Proc. ICASSP'02*, vol. I, pp. 701-704, 2002.

[10] C. L. Tsang, M. W. Mak, and S. Y. Kung, "Divergence-based out-of-class rejection for telephone handset identification," *Proc. ICSLP'02*, pp. 2329-2332, 2002.

[11] C. Mokbel, D. Jouvet, and J. Monne, "Deconvolution of telephone line effects for speech recognition," *Speech Communication*, vol. 19, pp. 185-196, 1996.

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

[13] 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.

[14] P. Färber, "Quicknet on MultiSpert: Fast Parallel Neural Network Training," *ICSI Technical Report TR-97-047, http://www.icsi.berkeley.edu/techreports/*, 1997.