Int J Speech Technol (2017) 20:753–759
DOI 10.1007/s10772-017-9447-8

Articulatory movement features for short-duration text-dependent speaker verification

Yan Zhang[2] · Yanhua Long[2] · Xiangrong Shen[1] · Haoran Wei[2] · Min Yang[2] · Hong Ye[2] · Hongwei Mao[2]

Received: 4 May 2017 / Accepted: 31 July 2017 / Published online: 7 August 2017

Abstract During our pronunciation process, the position and movement properties of articulators such as tongue, jaw, lips, etc are mainly captured by the articulatory movement features (AMFs). This paper investigates to use the AMFs for short-duration text-dependent speaker verification. The AMFs can characterize the relative motion trajectory of articulators of individual speakers directly, which is rarely affected by the external environment. Therefore, we expect that, the AMFs are superior to the traditional acoustic features, such as mel-frequency cepstral coefficients (MFCC), to characterize the speaker identity differences between speakers. The speaker similarity scores measured by the dynamic time warping (DTW) algorithm are used to make the speaker verification decisions. Experimental results show that the AMFs can bring significant performance gains over the traditional MFCC features for short-duration text-dependent speaker verification task.

Keywords Articulatory movement features · Dynamic time warping · Text-dependent · Speaker verification

✉ Yanhua Long
yanhua@shnu.edu.cn

Yan Zhang
aneybaby727@163.com

Xiangrong Shen
sxr@shnu.edu.cn

Haoran Wei
haoranwei@foxmail.com

Min Yang
yangmin@shnu.edu.cn

Hong Ye
yeeho@shnu.edu.cn

Hongwei Mao
hongweimao@shnu.edu.cn

[1] College of Humanities and Communications, Shanghai Normal University, Shanghai 200234, China

[2] Department of Electronical and Information Engineering, Shanghai Normal University, Shanghai 200234, China

1 Introduction

Speaker verification is a technique to automatically verify the speaker's identity based on the speaker's physiological and behavioral characteristics and features. In recent years, speaker verification technologies have reached an early level of maturity and have been widely deployed in commercial applications. Generally, speaker verification systems can be divided into text-dependent and text-independent ones. Text-dependent speaker verification requires users to pronounce according to a particular or fixed utterance text, training and test corpus must be text consistent. This type of speaker verification systems can normally achieve much better performances than the text-independent ones, given that the enrollment speech data is sufficient enough. However, they need users to cooperate because of the fixed texts. In text-independent systems, there are no constraints on the words which the speakers are allowed to use (Kinnunen and Li 2010). The training and the test utterances may have completely different content, which makes the text-independent speaker verification tasks much more challenging. However, because the enrollment and testing sessions are normally extremely short, text-dependent speaker recognition technology is particularly well suited for deployment in large-scale commercial applications (Hébert 2008). In this paper, we also focus on the short-time text-dependent speaker verification, the effectiveness of the proposed AMF features will be investigated.

The exploration of an effective front-end feature extraction which can capture the intrinsic characteristic of the individual speaker plays an important role in speaker verification task. The traditional acoustic features based on the spectral analysis, such as mel-frequency cepstral coefficients (MFCCs), linear predictive coding cepstral coefficients (LPCCs) (Young et al. 2002) and perceptual linear prediction coefficients (PLPs). However, these features are borrowed from speech recognition and they are not the best ones for speaker verification, especially for the short-time text-dependent verification systems, since these features are very easily affected by background environments or noises. Therefore, in recent years, more and more research efforts start to emphasize the contribution of looking for new features (Fu et al. 2014; Ganapathy et al. 2011).

To improve the speaker verification performances in reverberant environments, Ganapathy et al. (2011) proposed a frequency domain linear prediction (FDLP) feature to estimate long-term envelopes of speech in narrow subbands. Long et al. (2011) proposed a new feature based on the spectral subband energy ratios (SSERs) to characterize the speaker identity information resides in short-duration utterances. Smiliar to the SSERs, Alam et al. (2015) proposed a method to combine the amplitude spectrum, phase spectrum, and joint amplitude-phase based front-ends at score level to incorporate the complementary information. In recent years, with the development of deep neural network (DNN), various distinctive features have been proposed for the DNN-based estimation algorithms, such as, Guo et al. (2016) introduced a subglottal acoustic features estimated by using a DNN-regression model to better the performances of short-duration speaker verification tasks; Fu et al. (2014) presented three different types of deep features extracted from DNN and used them in a Tandem fashion to improve the text-dependent speaker verification; while Qian et al. (2016) investigated to use the bottleneck features and multilingual DNNs to narrow the gap caused by the data mismatch of i-vector system and DNN to improve the system performances. All of the features mentioned above do bring a better speaker recognition system. However, all of them are extracted still acoustic based features, they still can't avoid the distortion from background environments and noises.

Different from the previous works, in this paper, we exploit the speaker identity information from another aspect. We know that the speech is generated from the interaction movement of our articulators. Since the physiological structure of the articulator is unique for each speaker, in this paper, we assume that the features derived from the articulatory movement would be more robust and stable than traditional spectral based acoustic features for recognizing the speaker identities. We call these features the articulatory movement feature (AMF). Compared with the traditional acoustic features, AMFs have two advantages: (1) the articulatory features may be acquired by capturing the relative movement position of articulators directly, they are not influenced by acoustic and environmental background noises. Therefore, the AMFs can be more noise-robust and stable for capturing speaker individual characteristics than other acoustic features. (2) Due to physical constraints, articulatory features evolve in a relatively slow and smooth way. Hence, they are well-suited for speaker modeling with statistical models (Ling et al. 2009). Actually, with these potentially beneficial properties, the AMFs have been widely used in speech synthesis to improve the synthesized speech quality in the recent years (Ling et al. 2009; Cai et al. 2012; Toda et al. 2004). However, in this paper, instead of speech synthesis, we aim to investigate the possibilities by applying these AMFs to improve the short-duration text-dependent speaker verification systems. We invested great amount of time, resources, and efforts, to achieve spectacular improvement of performance, when compared to the traditional speech-based MFCC, on our own database, using some very old and very simple utterance matching technique. The dynamic time warping (DTW) algorithm is used to measure the AMF similarities between different speakers. Experiments are conducted on the 4-s training and 4-s testing verification task to validate the effectiveness of the proposed features.

The remainder of this paper is organized as follows. Section 2 presents the brief introduction to articulatory movement feature extraction. Section 3 introduces the speaker modeling frameworks. Section 4 demonstrates the experimental details and analysis of results. This paper is summarized in Sect. 5.

2 Articulatory movement feature extraction

Electromagnetic articulography (EMA) (Schönle et al. 1987) has the advantages such as convenient, accurate, real-time, compared with X-ray microbeam cinematography (Kiritani 1986), magnetic resonance imaging (MRI) (Baer et al. 1987), ultrasound (Akgul et al. 1998), and video motion capture of the external articulators (Summerfield 1987) etc. Therefore, similar to the AMF feature extraction in Ling et al. (2009) and Cai et al. (2012) for speech synthesis, we use EMA to collect AMF features. In this paper, the NDI wave equipment is used to record the articulatory movement positions and corresponding speech segments. As NDI wave uses EMA to collect AMFs, the articulatory movement features have been also called EMA parameters in some speech synthesis papers.

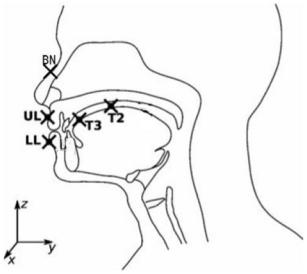

Fig. 1 Placement of the five EMA sensors

Table 1 Position labels for EMA sensors

Label	Position	Label	Position
T2	Middle tongue	UL	Upper lip
T3	Tip tongue	LL	Lower lip
BN	Nose bridge		

2.1 Feature extraction

To acquire the movement positions, four important positions for articulators and nose bridge are chosen to place sensors for our AMFs extraction. The placement positions of these sensors are shown in Fig. 1, and the labels are shown in Table 1. Each sensor records its space locatiossn of the corresponding articulators during the articulatory process. For each sensor receiver, coordinates in three dimensions spaces were recorded, the x axis represents from left to right of articulators, y axis represents from front to back of articulators, and z axis represents from bottom to top of articulators (Cai et al. 2012). Therefore, for each speech frame, we can obtain 15 position values from the five sensors with each sensor records 3-dimensional coordinate values. All five sensors were placed in the mid-sagittal plane of the speaker's head.

The 15 position values recorded by NDI wave are the original coordinate position values, it is not the best way to use them directly for speaker verification tasks, because the articulatory property of each individual speaker is reflected by the relative movement information of articulators to avoid the interference caused by head shaking of speakers. Hence, the original coordinate position values must be normalized. As the nose bridge and upper lip always remain relatively stationary with the head of the speaker, we take the nose bridge or upper lip as reference points, the relative positions of the low lip, tip tongue, middle tongue to the reference point formed the final AMF features used in this paper. They can be formulated as Eq. (1):

$$g(\nabla x_i, \nabla y_i, \nabla z_i) = a(x_i, y_i, z_i) - r(x_0, y_0, z_0) \quad (1)$$

where $a(x_i, y_i, z_i)$ represents the three-dimensional coordinates of the low lip, tip tongue, middle tongue, $r(x_0, y_0, z_0)$ represents the three-dimensional coordinates of reference points, and $g(\nabla x_i, \nabla y_i, \nabla z_i)$ represents the AMF features.

2.2 Database

We designed a short-duration text-dependent speaker verification database to validate the effectiveness of the proposed AMFs. Six speakers in total are included, with three male and three female speakers. The articulatory position values and its corresponding speech file are produced simultaneously during NDI wave recording. The range of these speakers ages are between 20 and 30 years old. Twenty-five different sentence-level texts are designed for speech recording. Each speaker read the 25 text utterances twice at different time. The length of each speech recording is around 4 s with 22.05 kHz sampling frequency. The articulatory position values are sampled at a frequency of 400 Hz.

During the recording of EMA parameters, it is normal to occur the burr phenomenon that, some EMA parameters of speech frames are not well collected by the NDI wave system because of the sensitivity of the sensors placed on our articulators. Therefore, we can not guarantee that all of the recorded speech recordings have a completely well-collected EMA parameters saved in the parameter file, the EMA parameters of some speech frames are missing. Some researchers used the parameter interpolation method to alleviate this problem (Tachibana et al. 2005), however, the accuracy of those interpolated parameters still need to be greatly improved. Therefore, in our experiments, we recorded four times for each text per speaker then selected the needed data files without missing EMA parameter to construct the training and testing speaker verification trials.

3 Speaker modeling

By using feature vectors extracted from a given speaker's training utterance(s), a speaker model is trained to represent the claimed speaker identity and stored into the system database for testing. For text-dependent speaker verification, the speaker model is normally utterance-specific and it includes the temporal dependencies between the feature vectors. The classical approach used to measure the degree of similarity between training and test feature vectors is the dynamic time warping (DTW) (Furui 1981) and the hidden markov model (HMM) (BenZeghiba and Bourland 2006; Naik et al. 1989). In this paper, we use the DTW for our short-duration text-dependent speaker verification tasks.

DTW is one kind of dynamic programming techniques (Furui 1981). Since the length of the feature vectors for

training and test utterances may vary in time or speed in speaker verification tasks. The unequal length of two series of feature vectors represents different speech rates of different speakers, even the same speaker may has different speech rates in different speech recordings. Therefore, before comparing the similarity between training and test feature vectors, one or two sequences need to be warped non-linearly along its time axis so as to find the corresponding regions between the two time series (Muda et al. 2010). In speaker verification, an overall cumulative distance between the test and the training feature vectors is obtained using the DTW, this cumulated distance is then used to compare with a threshold to determine whether to accept or reject an identity claim.

Suppose we have two sequences of feature vectors Q and C, whose lengths are n and m respectively:

$$\begin{aligned} Q &= q_1, q_2, \ldots, q_i, \ldots, q_n; \\ C &= c_1, c_2, \ldots, c_j, \ldots, c_m; \end{aligned} \tag{2}$$

In order to align these two sequences with DTW algorithm, an $n \times m$ matrix grid need to be constructed, matrix elements (i_{th}, j_{th}) denotes the alignment distance $d(q_i, c_j)$ between the two feature vectors q_i and c_j. The similarity between each pair of sequence Q and C is that, the smaller the distance, the higher the similar degrees. The Euclidean distance is used to calculate the distance between two feature vectors:

$$d(q_i, c_j) = (q_i - c_j)^2 \tag{3}$$

Then, the accumulated distance is measured by Eq. (4):

$$D(i, j) = \min[D(i-1, j-1),\ D(i-1, j), D(i, j-1)] + d(q_i, c_j) \tag{4}$$

4 Experiments

4.1 System configuration

The database used for our experiments is recorded by NDI wave. This database consists two parts: the first part is EMA feature parameters collected by five sensors, the second part is the speech audio files recorded by microphone. The detail to obtain the speech and EMA parameters, the AMF feature pre-processing methods can be referred to Sect. 2. Three-dimensional feature is used. In our experiments, 150 target speaker enrollment utterances and 150 test utterances are included, 150 speaker models are build. In total, we have 150 target speaker trials and 750 imposter speaker trials for the system evaluation.

4.2 Speaker discrimination in feature space

4.2.1 AMF discrimination

Figure 2 illustrates the discrimination in the AMF feature spaces between two different speakers with different (female speaker-spkA and male speaker-spkB) and the same speaker gender information (male speakers, spkB and spkC). These AMFs are the 3rd dimension, which represents the articulators movement trajectory relative to the low lip reference points of tip tongue (a, d), middle tongue (b), and low lip (c) respectively. All of the speakers said the same sentence "7点40 喊我起床 (wake me up at 7:40)" in Chinese. It can be seen from the first three subfigures (a–c) that, the trajectories of the same dimension of AMFs are similar for both speakers, because they said the same content. However, when comparing the amplitude values, it is easy to note the big difference of AMFs in the ∇Z_i when they speaking, because the physiological and behavioral characteristics are different for different speakers even they say the same sentence. Furthermore, when comparing the subfigure (a) and (d), we can observe that, even spkB and spkC are the same gender speakers, the AMFs' difference between them is still big. Therefore, motivated by the observations from these amplitude differences between different speakers, we expect that the AMF features may provide better discrimination than other traditional acoustic feature to build a speaker recognition system.

4.2.2 MFCC discrimination

In order to see the traditional acoustic feature discrimination between two speakers, we take the MFCC C0 as an example. The distribution of MFCC C0 and its first and second order derivatives from the female and male speakers (spkA, spkB) are shown in Fig. 3. The same as the AMFs in Fig. 2, both of the two speakers said the same sentence "7点40喊我起床 (wake me up at 7:40)" in Chinese. Different from the big discrimination we observed from Fig. 2, the trajectories and the distributions of MFCC C0 between two speakers are very similar, even these two speakers with totally different gender information.

4.3 Baseline system

We take the results from the traditional 39-dimensional MFCC features (13 dimensional MFCCs and their first and second order derivatives) as our baseline. As presented in Sect. 3, the DTW algorithm is used to obtain the speaker identity similarities between the train and testing speech utterances for both the MFCCs and the AMFs.

We use the equal error rate (EER) as the evaluation metric to examine the system performances, and an

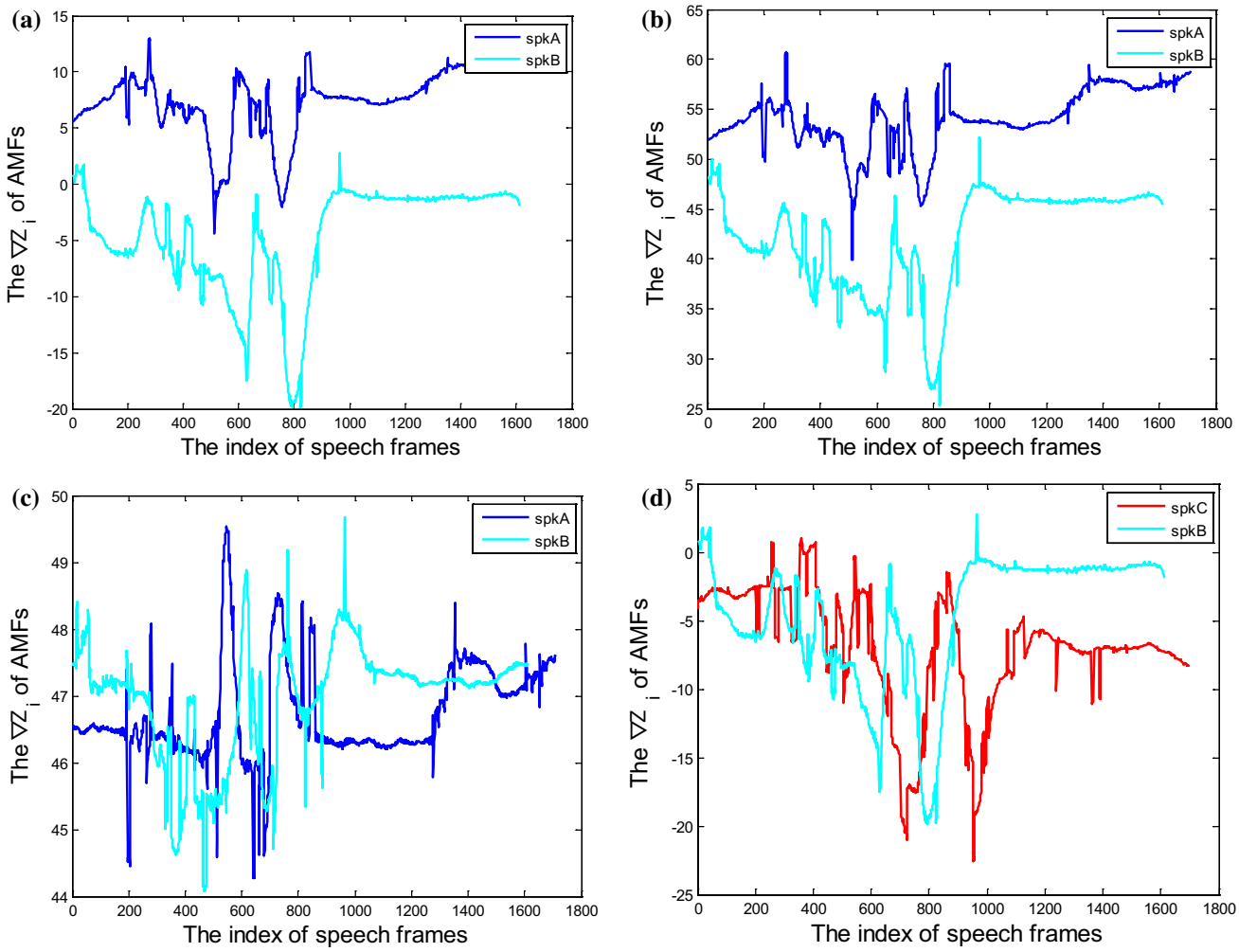

Fig. 2 Illustration of the AMFs' discrimination for different speakers. **a** The 3rd dimension of AMFs (collected from *tip* tongue). **b** The 3rd dimension of AMFs (collected from *middle* tongue). **c** The 3rd dimension of AMFs (collected from *low* lip). **d** The 3rd dimension of AMFs (collected from *tip* tongue) for two male speakers

EER = 6.27% is obtained from the baseline system. The performances derived from the AMFs will be compared with this baseline EER to show their effectiveness for speaker verification.

4.4 AMFs evaluation

Table 2 illustrates the EER performances derived from the AMFs derived from the low lip, tip tongue, and middle tongue respectively.

From Table 2, it's clear to see that, the reference point plays an important role during the collecting of AMF features. There are big EER gaps between the AMFs collected from the low lip, tip tongue and middle tongue by using the nose bridge and up lip as reference points. Moreover, the behaviors of AMFs derived from different articulators are also different. Compared with using the nose bridge as reference point, the AMFs from low lip and tip tongue achieved much lower EERs when using the up lip as the reference point. However, the behavior of AMFs derived from the middle tongue goes in the opposite direction.

In addition, compared with the baseline with MFCC features, the AMFs achieved much better or similar performances with different articulators by taking the up lip as reference point. Therefore, we expect that, the combination of all the three articulators will give us the best results, because in the common sense, we think that, the interaction movement of all the articulators is more complex than the single articulator, the combination of AMFs from all of the three articulators may provide stronger discrimination of the identity for each speaker than any of the AMFs shown in Table 2.

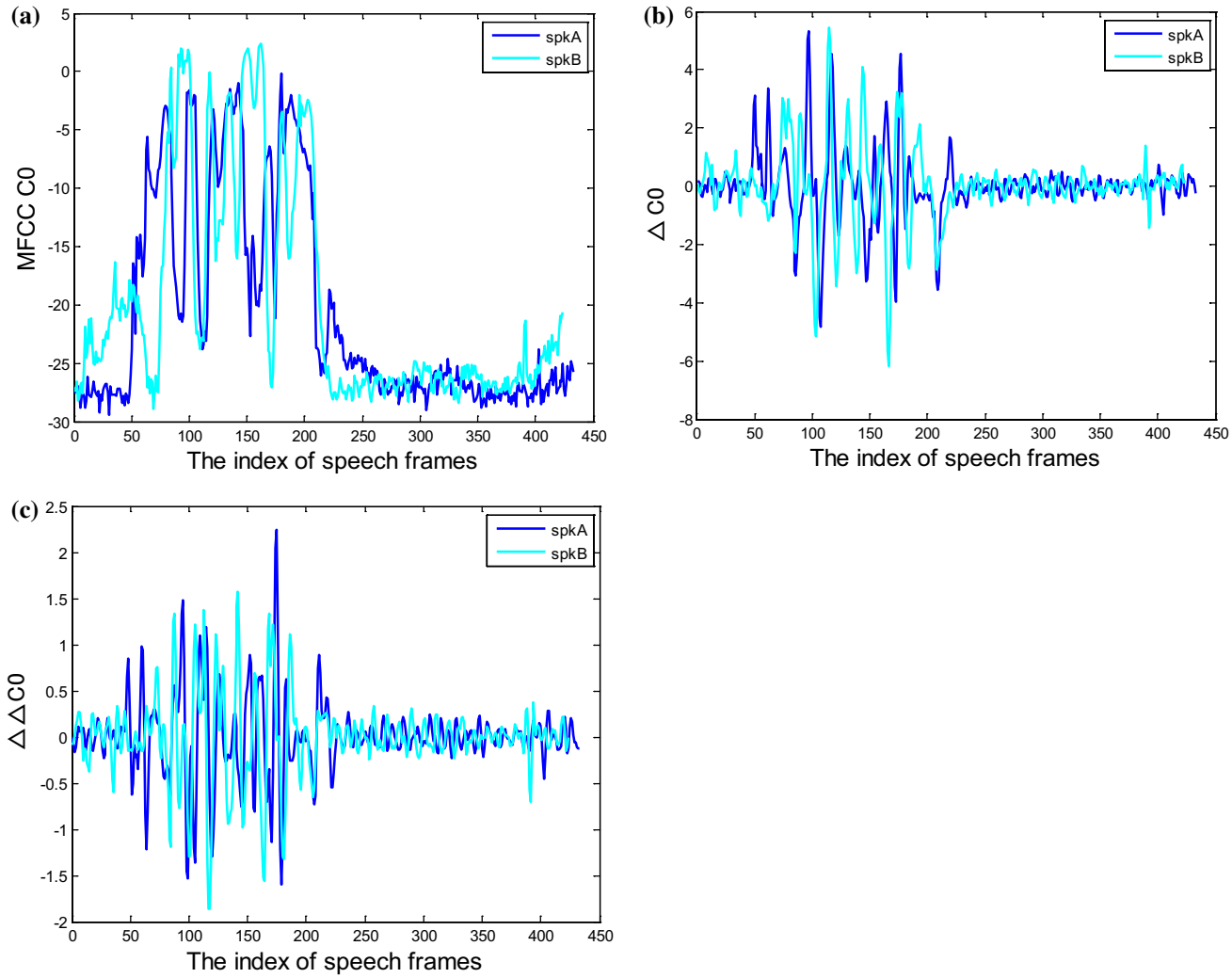

Fig. 3 The distribution of MFCC C0 comparison for different speakers. **a** Distribution of C0. **b** Distribution of ΔC0. **c** Distribution of ΔΔC0

Table 2 EER% with different AMFs

Reference point	Articulators		
	Low lip	Tip tongue	Middle tongue
Nose bridge	13.6	11.5	1.8
Up lip	2.1	4.7	6.8

Table 3 EER% after feature and system fusion

	Feature fusion	System fusion
EER	1.53%	0.49%

4.5 Feature and system fusion

The effectiveness of combining the AMF features both in the feature and system level is examined in this section.

In the feature combination level, we just directly concatenate each 3-dimensional AMFs derived from the low lip, tip tongue and middle tongue to form a 9-dimensional AMF feature vector, taking the upper lip as their reference point during AMF feature extraction. These 9-dimensional AMFs are then used as the input features for DTW-based speaker verification. However, in the system fusion level, we train three speaker verification systems independently on each of the 3-dimensional AMFs which are used to form the 9-dimensional AMF feature vector, then three of the DTW similarity scores are average as the final score for each test trial to make the accept or reject decision. Results are shown in Table 3.

According to Table 3, it is clear to obtain that both the feature and system fusion work very well for speaker verification and give significant EER reduction than the numbers in Table 2. This tell us that the interaction movement of different articulators do provide more discriminative information than single one, it has a better speaker identity characterization ability. In addition, we can see that the system fusion at the score level is superior to the combination at the AMF feature level, a relative 68% improvement is obtained, reducing the EER from 1.53 to 0.49%.

5 Conclusion

In this paper, a new feature of AMFs is proposed for short-time text-dependent speaker verification. The idea is borrowed from the articulator movement parameters used in speech synthesis. Experimental results show that the proposed AMFs work very well, they can provide much stronger discriminative information to distinguish different speakers than the conventional MFCC features for speaker verification. Although the AMFs are proved to be effective, at present, the high cost and scale of effort needed for implementing EMA is still a disadvantage, which does not facilitate the practical use of such technology, however, this research provides a new perspective on using EMA parameters to distinguish speaker identities, which is away from the mainstream research. In our future research, we will focus on exploring effective method to solve the burr problem to improve the AMF feature extraction, applying the AMFs to text-independent speaker verification tasks to see its effectiveness, and investigating some state-of-the-art machine learning techniques to model the AMFs more effectively.

Acknowledgements This work was funded by the Shanghai Normal University (Grant No. DCL201702) and Shanhai Sailing Program, Science and Technology Commission of Shanghai Municipality (Grant No. 14YF1409300).

References

Akgul, Y. S., Kambhamettu, C., & Stone, M. (1998). Extraction and tracking of the tongue surface from ultrasound image sequences. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 298–303).

Alam, M. J., Kenny, P., & Stafylakis, T. (2015). Combining amplitude and phase-based features for speaker verification with short duration utterances. In *Proceedings of the Interspeech* (pp. 249–253).

Baer, T., Gore, J. C., Boyce, S., et al. (1987). Application of MRI to the analysis of speech production. *Magnetic Resonance Imaging, 5*(1), 1–7.

BenZeghiba, M., & Bourland, H. (2006). User-customized password speaker verification using multiple reference and background models. *Speech Communication, 48*(9), 1200–1213.

Cai, M. Q., Ling, Z. H., & Dai, L. R. (2012). Target-filtering model based articulatory movement prediction for articulatory control of HMM-based speech synthesis. In *Proceedings of the IEEE International Conference on Signal Processing* (pp. 605–608).

Fu, T., Qian, Y., Liu, Y., & Yu, K. (2014). Tandem deep features for text-dependent speaker verification. In *Proceedings of the Interspeech* (pp. 1327–1331).

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing, 29*(2), 254–272.

Ganapathy, S., Pelecanos, J., & Omar, M. K. (2011). Feature normalization for speaker verification in room reverberation. In *Proceedings of the ICASSP* (pp. 4836–4839).

Guo, J., Yeung, G., Muralidharan, D., et al. (2016). Speaker verification using short utterances with DNN-based estimation of subglottal acoustic features. In *Proceedings of the Interspeech* (pp. 2219–2222).

Hébert, M. (2008). Text-dependent speaker recognition. In *Springer handbook of speech processing*. Berlin: Springer.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication, 52*(1), 12–40.

Kiritani, S. (1986). X-ray microbeam method for measurement of articulatory dynamics-techniques and results. *Speech Communication, 5*(2), 119–140.

Ling, Z. H., Richmond, K., Yamagishi, J., & Wang, R. H. (2009). Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio Speech and Language Processing, 17*(6), 1171–1185.

Long, Y., Yan, Z. J., Soong, F. K., Dai, L., & Guo, W. (2011). Speaker characterization using spectral subband energy ratio based on harmonic plus noise model. In *Proceedings of the ICASSP* (pp. 4520–4523).

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing, 2*(3), 138–143.

Naik, J., Netsch, L., & Doddington, G. (1989). Speaker verification over long distance telephone lines. In *Proceedings of the ICASSP* (pp. 524–527).

Qian, Y., Tao, J., Suendermann-Oeft, D., Evanini, K., & Ivanov, A. V. (2016). Noise and metadata sensitive bottleneck features for improving speaker recognition with non-native speech input. In *Proceedings of the Interspeech* (pp. 3648–3652).

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language, 31*(1), 26–35.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio–visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). Hove, UK: Lawrence Earlbaum Associates.

Tachibana, M., Yamagishi, J., Masuko, T., & Kobayashi, T. (2005). Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions on Information and Systems, 88*(11), 2484–2491.

Toda, T., Black, A. W., & Tokuda, K. (2004). Mapping from articulatory movements to vocal tract spectrum with gaussian mixture model for articulatory speech synthesis. In *5th ISCA Speech Synthesis Workshop* (pp. 31–36).

Young, S., Evermann, G., & Gales, M. J. F. (2002). *The HTK book*. Cambridge: Cambridge University Engineering Department.