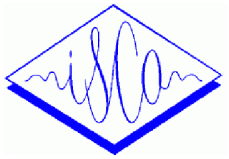
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Cross Linguistic Comparison of Mandarin and English EMA Articulatory Data

Sheng Li[1,2], *Lan Wang*[1,2]

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2] The Chinese University of Hong Kong, Hong Kong, China
{sheng.li, lan.wang}@siat.ac.cn

Abstract

This paper aims at effectively identifying common English mispronunciations by Mandarin speakers and incorporating this knowledge into computer assisted language learning (CALL) to improve the learner's accented English. For this purpose, English and Mandarin multi-channel EMA articulatory datasets collected from native English and native Mandarin speakers respectively have been used to uncover cross-linguistic distinctions. The Procrustes based speaker normalization is used to eliminate the variability which comes from speaker-specific vocal-tract anatomies and other individual biomechanical properties. Then the English phonemes missing from Mandarin and their Mandarin confusing equivalents are identified using phonological knowledge. These English and Mandarin phoneme pairs may be hard to distinguish in acoustics, but by extracting useful information from the changing on tongue positions and shapes of the lips while speaking can be good cross linguistic phoneme level comparison metrics both empirical and quantified. With this method, the same analysis can be done between languages, or different accents within the same language in the future.

Index Terms: mispronunciation, articulatory data, cross linguistic comparison

1. Introduction

Recent development of computer assisted pronunciation training (CAPT) has benefit a lot from current automatic speech recognition (ASR) and speech visualization techniques. It also gets direct instruction from the explorations on speech production and perception. These linguistic researches have been no longer relying on auditory analysis, but also on measuring the activities of the articulators (the tongue, the larynx, the lips and the jaw) during speech. Many devices, such as X-ray microbeam cinematography, Cine-MRI, ultrasound, electropalatography and electromagnetic articulography (EMA) are used for this purpose.

According to the theory of language transfer **[1]**, it is assumed that the learner's mother tongue may negatively affect his learning a foreign language. Such effects were observed when analyzing the mispronunciations made by Chinese learners. We find that the English phonemes which are missing from Mandarin may most easily be mispronounced or even replaced by Mandarin phonemes.

The objective of our research in this paper is to find those English phonemes, which may most probably lead to confusions and mispronunciations by Mandarin speakers, and their equivalents in Mandarin, so that we can incorporate this knowledge into the CALL system.

For this purpose, we had collected English and Mandarin multi-channel EMA articulatory data collected from native English and native Mandarin speakers respectively. These datasets provide us a chance to uncover cross-linguistic distinctions. But the major challenge here is how to overcome the variability that comes from speaker-specific vocal-tract anatomies and other individual biomechanical properties.

Current cross linguistic articulatory study can be found in the experimental phonetics researches. In **[2]**, German and Hungarian tongue shape comparisons of articulatory profiles were carried out on both static and kinematic tongue configurations. The work of **[3]** summarized the techniques for speaker normalization derived from Procrustes methods **[4]** could be effectively applied to both acoustic and articulatory data.

The other studies related to articulatory speaker normalization are the series researches about Audio–Visual Speech Processing (AVSP). These researches concern more about constructing a speaker-independent statistical model (GMM, HMM and etc.) like in **[5]**, coped with speaker adaptation techniques. These methods have been well developed in speech recognition, but required a large scale of multi-speaker datasets.

For quantified comparison of the ariticulatory data, the method of projecting the phonemes onto a universal articulatory space was investigated in **[6]**, which used multi-dimensional scaling (MDS) algorithm **[7]**. Alternatively, the research in **[8]** also introduced Hierarchical Clustering Analysis (HCA) **[9]** to generate the classes of the equivalent phonemes.

The methodology we choose is as follows: we use the Procrustes based speaker normalization just considering our experimental condition of limited data. Then we identify the English phonemes missing from Mandarin and their equivalent in Mandarin and give empirical comparison. For quantified comparison, we visualize distances of all the phonemes from two languages onto a quantitative and cross-linguistic phonetic space by multi-dimensional scaling (MDS) analysis. Hierarchical Clustering Analysis (HCA) also has been used to cluster the similar phonemes from two different languages.

The rest of this paper is organized as follows: Section 2 introduces the collection and data processing of EMA data. Section 3 describes what we do to normalize the speaker difference between the two language data. In Section 4, the normalized data is used to construct a cross linguistic and speaker independent articulatory space, so that mispronunciation confusions can be observed directly. The conclusions and future work are in Section 6.

2. Recording and Processing the Multichannel EMA Articulatory Data

2.1. Corpus Design and Speakers

In our previous research **[10]**, we had recorded an English EMA dataset from a native English speaker for the purpose of synthesizing dynamic phoneme level articulation and drive a 3D animation to speak. Now we have recorded yet another Mandarin dataset from a native Mandarin speaker so as to build a bilingual multi-channel EMA articulatory database. The corpus and speaker information for these two datasets are listed as follows.

Table1. Corpora design

	Speaker	Prompts	Dur.
English Data	Female, Age 22, Florida, USA	45 phones, 125 word pairs, 1 short passage	20min
Mandarin Data	Female, Age 40, Beijing, China	21 initials and 35 finals 23 word pairs, 6 tongue-twisters, 1 short passage	25min

2.2. Data Collection

In the data collection, the facial and intra-oral articulator movements are recorded using the Carstens EMA AG500 at a sample rate of 200 frames per second, and an accuracy of motion tracking approximately at 0.5 mm **[11]**.

The movements are synchronized to the speech waveform. As in previous studies **[12]**, 3 coils are placed on the speaker's lips, 1 coil on her jaw, 3 coils on her tongue to record the internal movements, and 3 coils on nose bridge and skull behind the ears.

2.3. Data Processing

Since the recorded sensor motion is usually a mixture of head movement and the actual articulator motion. The head motion normalization is required to eliminate the head movement from the recorded data. So 3 coils (H1, H2, H3) on Nose Bridge and skull behind the ears are used for head motion normalization **[13]**.

We did down sample rate to the EMA data for a smoother trajectory. When processing the audio data in ASR tasks, we also labeled the multichannel EMA data synchronized with audio data.

To define the characteristics of speech production and also for data reduction, we borrow the conception of 'key frame' and 'static frame' for each phoneme as in our previous work **[14] [15]**. The static frame is selected from the data in relaxing state to define the starting point of each articulatory movement. The key frame is defined as the peak position of an individual phoneme, which should refer to the characteristics of producing this sound. For instance, the peak position of the phoneme /ɛ/ should be selected with the maximally opened mouth, while the tongue is also at its lowest point. Fig.1 depicts the lips and tongue positions of individual vowel /ɛ/. It shows the profile view of lips and tongue at the feature state, in contrast to that of the static state.

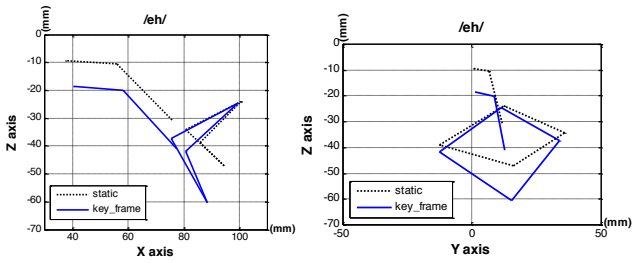


Figure.1 Key frame and static frame of /ɛ/

3. Techniques for Comparing Cross Linguistic Articulation

3.1. Contrastive Analysis

We first use the contrastive analysis to identity the English phonemes missing from Mandarin. We use General American English phonemic inventory based on **[16]** and Standard Mandarin phonemic inventory based on **[17]**.

Table2 lists the English phonemes missing from Mandarin and their Mandarin equivalents. In this table, Phonemes in American English and Mandarin Chinese are all given in form of International Phonetic Alphabet (IPA) symbols. And Mandarin phonemes have additional notations of Chinese characters.

Table2. Main English phonemes missing from Mandarin and their Mandarin equivalences [18]

	English phonemes missing from Mandarin	Mandarin equivalents
Affricates	post-alveolar affricate /dʒ/	/tʃ/(知)
Fricatives	inter-dental fricative /θ/	/s/(思)
	inter-dental fricative /ð/	/z/(资)
	voiced fricative /v/	/f/(佛)
	voiced fricative /z/	/ts/(资)
	voiced fricative /ʒ/	/tʃ/ (知)
High vowels	lax vowels /ʊ/	/u/(乌)
	lax vowels /ɪ/	/i/(衣)
Mid vowels	mid-low front vowel /ɛ/	/ai/(挨)
	rounded mid-low back vowel /ɔ/	/o/(喔)
Low vowels	low front vowel /æ/	/ai/(挨)
	low central vowel /ʌ/	/a/(啊)

3.2. Speaker Normalization

To overcome the speaker variability, we investigated the Procrustes transformation **[4]** is a linear geometric transform from the source multi-point object to the target multi-point object according to a least mean square criterion. This method includes a global transform, scaling, and rotation which preserve the relative anatomical shape of the source speaker.

Suppose we seek to develop a transformation that normalizes articulatory data X_1 of the source speaker to the articulatory data X_2 of the target speaker, the normalized

articulatory data of X_1 is X_3. The transform can be expressed in Equation 1. as a combination of rotation (H), scaling (b) and translation (c).

$$X_3 = H X_1 b + c \quad (1)$$

The rotation matrices H are calculated by performing singular value decomposition (SVD).

$$(X_1')^T X_2' = U \Sigma V^T \quad (2)$$

$$H = VSU^T \quad (3)$$

where Σ is diagonal matrix and U,V are orthogonal. And S is a diagonal matrix with $|s_{ii}| = 1$, and the signs of the s_{ii} are taken from the corresponding elements of Σ. The normalization parameters (H, b, c) are optimized by the least-squares criterion applied between the target and normalized source articulators.

3.3. Comparing Cross Linguistic Articulatory Motions

In our previous work, we also did preliminary tongue position analysis for 3D articulation animation **[10]**. We used the clustering of the phoneme level displacements to figure out the linguistic distinctiveness of major English vowels.

The evaluation can be taken out in two forms, one is subjective, and the other is objective. In subjective evaluation, we can visualize the articulation of the confusable phone pairs in a normalized articulatory space.

In objective evaluation, we can evaluate the cross linguistic distinctiveness by computing the Mahalanobis distances of displacements of the articulators (3 points on the tongue, and 3 points on the lips). With these pair wise distances, we can derive a quantitative vowel space as the perceptual dissimilarity matrices using multi-dimensional scaling (MDS) analysis. Hierarchical Clustering Analysis (HCA) can also be used here to generate the classes of the equivalent phonemes.

4. Experiments and Discussions

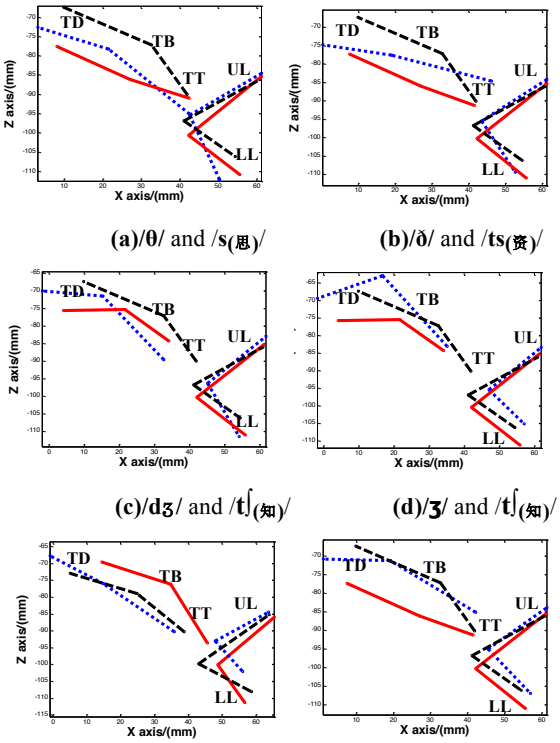
The English and Mandarin phonemes listed in table2 may be similar in acoustics, but they will absolutely not have the same tongue positions and shapes of the lips while speaking. So these phoneme pairs are hard to distinguish just by listening, but their differences are obvious from the visualized articulatory data.

Figure 2 and Figure 3 listed empirical comparisons taken in different groups between these easily confused cross linguistic phone pairs. Figure 2 and Figure 3 label the tongue tip as TT, tongue body as TB, tongue dorsum as TD, upper lip as UL and lower lip as LL. (The solid/red line is the Mandarin key frame, the dash/black line is English key frame, and the dot/blue line is the static frame.)

The listed English consonants missing from Mandarin are mainly fricatives and affricates, which can be categorized according to different points of articulation.

/θ/ and **/ð/** are inter-dental fricative and their equivalents in Mandarin **/s(思)/** and **/ts(资)/**, whose points of articulation moved backward to alveolar. The other groups (**/z/** and **/ts(资)/** are alveolar, /ʒ/, **/dʒ/**, **/tʃ(知)/** are post-alveolar, **/v/**, **/f(佛)/** are labiodental) have the same points of articulation, but the differences mainly come from the shapes of lips opening and tongue position.

Many Chinese students can't distinguish the lax vowels and the tense vowels. They would think the differences mainly come

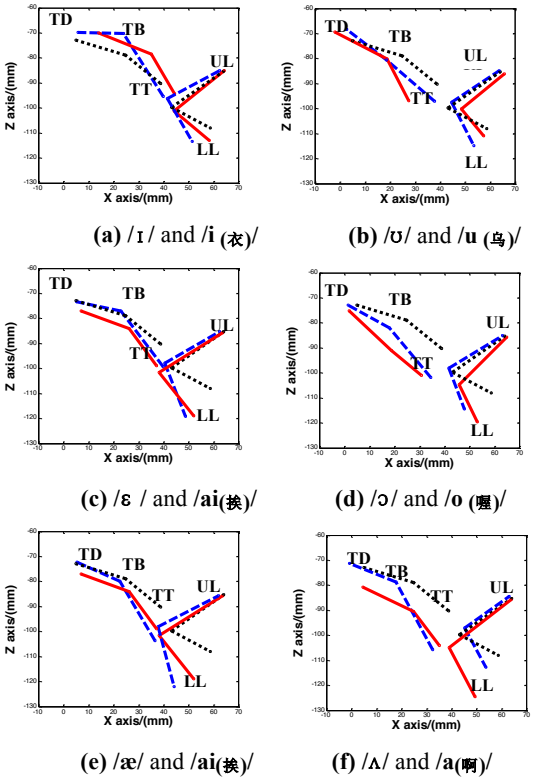


(a)/θ/ and **/s(思)/** **(b)/ð/** and **/ts(资)/**

(c)/dʒ/ and **/tʃ(知)/** **(d)/ʒ/** and **/tʃ(知)/**

(e)/v/ and **/f(佛)/** **(f) /z/** and **/ts(资)/**

Figure 2.English consonants missing from Mandarin and their Mandarin equivalents



(a) /ɪ/ and **/i (衣)/** **(b)** /ʊ/ and **/u (乌)/**

(c) /ɛ/ and **/ai(挨)/** **(d)** /ɔ/ and **/o (喔)/**

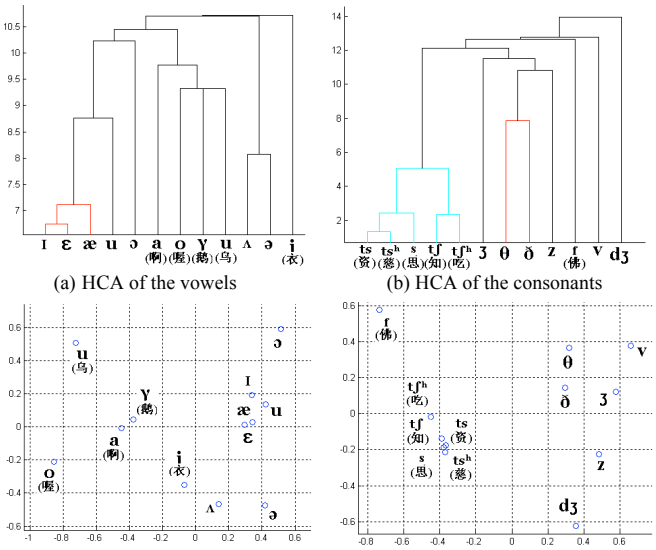
(e) /æ/ and **/ai(挨)/** **(f)** /ʌ/ and **/a(啊)/**

Figure3. Tongue positions and lip opening of English vowels missing from Mandarin and their Mandarin equivalents

from the durations. They actually neglect that English vowel system has more complex tongue position classification. For other vowels like **/ɛ/** and **/æ/,** students can't find their equivalent in Mandarin.

To objectively evaluate the dissimilarities of the across the linguistic phoneme pairs, we can compute the Mahalanobis distances between their displacements of the articulators (3 points on the tongue, and 3 points on the lips).

Hierarchical clustering analysis (HCA) and multi-dimensional scaling (MDS) analysis are all effective ways to visualizing the distinctiveness of the cross linguistic phonemes.



(a) HCA of the vowels (b) HCA of the consonants

(c) MDS analysis of the vowels (d) MDS analysis of the consonants

Figure.4 visualizing the cross linguistic phoneme distinctions in two ways

In Fig4. the results of both methods can support each other. The dissimilarity information showed from the 2-D MDS structures and the clustering tree structures. Fig4 suggest that although the phoneme pairs sound like each other, the phoneme level articulations of Mandarin are significantly different from those of English.

5. Conclusions and Future Work

Our research aims at effectively identifying common English mispronunciations by Mandarin speakers and incorporating this knowledge into CALL system to improve the learner's accented English.

We had English and Mandarin multi-channel EMA articulatory data collected from native English and native Mandarin speakers respectively. These datasets provide us a chance to uncover cross-linguistic distinctions.

We choose the Procrustes based speaker normalization to eliminate the variability which comes from speaker-specific vocal-tract anatomies and other individual biomechanical properties. Then we identity the English phonemes missing from Mandarin and give both empirical and quantified comparison metric on the cross linguistic phoneme level articulatory data.

With this method, someday in the future we can do the same analysis between languages, or different accents within the same language and incorporate this knowledge into our future language learning software both for 2nd language learners and hearing impaired children.

6. Acknowledgements

Our work is supported by National Nature Science Foundation of China (NSFC 61135003, NSFC 90920002), and The Knowledge Innovation Program of the Chinese Academy of Sciences (KJCXZ-YW-617).

7. References

[1] R. Ellis. "*The Study of Second Language Acquisition*", Oxford University Press, 1994.

[2] C. Geng, "*A cross-linguistic study on the phonetics of dorsal obstruents*", Doctoral Dissertation, 2008

[3] C. Geng and C.Mooshammer. "*How to stretch and shrink vowel systems: results from a vowel normalization procedure*",Journal of the Acoustical Society of America 125(5): 3278-3288, 2009.

[4] J. C. Gower, "*Generalized Procrustes analysis*", Psychometrika 40, 33–51, 1975

[5] S. Hiroya, and T. Mochida, ``*Multi-speaker articulatory reconstruction based on an eigen articulatory HMM*," in Proc. ICASSP, pp. 909-912, March 2005.

[6] J.Wang, J. R. Green, A. Samal and D. B. Marx, "*Quantifying articulatory distinctiveness of vowels*", in Proc InterSpeech, pp.277-280, Florence, Italy, 2011.

[7] J.B. Kruskal and M. Wish, "*Multidimensional scaling*". Beverly Hills, CA and London: Sage Publications, 1978.

[8] J. Jiang, "*Relating optical speech to speech acoustics and visual speech perception*", Doctoral Dissertation, University of California, Los Angeles, 2003

[9] M.S. Aldenderfer and R.K. Blashfield, "*Cluster analysis*". Beverly Hills and London: Sage Pubns, 1984.

[10] S. Li, L.Wang and E. Qi, "*The Phoneme-Level Articulator Dynamics for Pronunciation Animation*", in Proc International Conference on Asian Language Processing (IALP), pp.283-286, 2011, Penang, Malaysia

[11] Y. Yunusova, J. R. Green, and A. Mefferd, "*Accuracyassessment for AG500 electromagnetic articulograph*", Journalof Speech, Language, and Hearing Research, vol. 52(2) pp.547-555, 2009.

[12] Y. Tarabalka, P. Badin, F. Elisei and G. Bailly, "*Can You Read Tongue Movements? Evaluation of The Contribution of Tongue Display to Speech Understanding*", in Proc ASSISTH 2007, France, pp. 187-190, November 2007.

[13] C. Kroos, "*Using sensor orientation information for computational head stabilisation in 3D Electromagnetic Articulography (EMA)*". In Proc InterSpeech 2009, pp.776-779, Brighton, UK, 2009

[14] L. Wang, H. Chen and J. Ouyang, "*Evaluation of External and Internal Articulator Dynamics for Pronunciation Learning*", in Proc Interspeech 2009. Pp. 2247-2250, UK, Sep. 2009.

[15] H. Chen, L. Wang, P.A. Heng, and W.X. Liu, "*Combined X-Ray and Facial Videos for Phoneme-level Articulator Dynamics*", The Visual Computer, 2010.

[16] P. Ladefoged, "*A Course in Phonetics, 4th ed"*. Boston: Heinle & Heinle. 2001

[17] W. Lee and E. Zee, "*Standard Chinese (Beijing).*" Journal of International Phonetic Association 33:1, 109-112, 2003

[18] A. Harrison, "*Contrastive analysis of Mandarin, Cantonese and English phonologies*", Interim Report, CUHK, 20 Jan 2008