# Towards a quantitative model of Mandarin Chinese perception of English consonants

*Jian Gong, Martin Cooke and M. Luisa García Lecumberri*
*University of the Basque Country, Spain, Ikerbasque (Basque Foundation for Science), Spain*

ABSTRACT

Models which describe the perception of foreign language sounds typically do so using qualitative relations with sounds in the first language, but a more fine-grained account of learners' perceptual difficulties might be obtained via techniques from automatic speech recognition. Here, we employ generative statistical models of speech – Hidden Markov Models – which underly most work in speech recognition, to learn the sound systems of English and Mandarin Chinese, and use these as the basis for a quantitative model of the perception of English intervocalic consonants by Chinese listeners. This approach allows both the prediction of consonant identification preferences and the construction of a complete cross-language consonant distance matrix, which can then be compared to consonant categorisation and goodness ratings respectively. To evaluate the model, 30 native Chinese listeners with moderate English competence and residing in China categorised and rated English intervocalic consonants. A high degree of consistency was found between listeners and the computer model for sound categorisation as well as a clear and significant correlation between goodness ratings and distance measurements, suggesting that the technique has potential in predicting sound-level difficulties experienced by language learners. One key difference between listeners and the model is in the relative influence of acoustic and orthographic factors in sound categorisation: while model decisions and distances are based solely on acoustic information, clear evidence of orthographic interference can be seen in listener responses. We explored this issue by comparing the use of Chinese characters versus Pinyin graphemes on the task. Listeners who were presented with Pinyin showed some orthographic influences, while a dialect influence occurred in the character group.

KEYWORDS: consonant categorisation, goodness, computer model

## 1. Introduction

Learners' L2 sound perception is strongly influenced by their L1 sound system. Theoretical models of second language acquisition have been proposed to give predictions of the degree of success in L2 sound acquisition. The Perceptual Assimilation Model (PAM) (Best 1995) suggests that the perception of L2 sounds is based on their similarities to or distances from the closest counterpart sounds in the L1. Accordingly, the perceived distance between L2 and L1 sounds may determine the degree of discriminability of certain L2 sound pairs. The Speech

Learning Model (SLM) (Flege 1995) hypothesises that only when the phonetic distance between the L2 sound and L1 sound is large enough for the learner to detect, a new phonetic category may be established. The larger the perceived distance between an L2 and an L1 sound, the more likely a new category for the L2 sound will be established during the learning process. Accordingly, both models agree that the perceived phonetic distance between L1 and L2 plays a crucial role in L2 perception. Many studies have been carried out to test the predictions made by the two models via cross-language mapping experiments and categorical discrimination tests (Strange et al. 1998, 2001; Guion et al. 2000; Lengeris and Hazan 2007). These studies show that although the two models have a certain degree of explanatory power, there are still some limitations. Since the PAM and SLM describe how L1-L2 sound similarity or distance affect non-native perception in a mainly qualitative manner, a better understanding of the relationship between the L1 and L2 sound systems might be obtained using a more quantitative approach. A recent study shows the benefit of using a statistical approach to predict Chinese L2 learners' perception of English vowels (Thomson et al. 2009). Current study's goal is to develop a computational account of L2 sound perception capable of predicting non-native confusions. Such a model would allow a controlled examination of the influence of factors such as the amount and quality of sound exposure as well as an exploration of how two or more co-existing sound systems interact in classifying speech sounds. Here we present an initial model of Chinese perception of English intervocalic consonants, and compare its responses with behavioural data from Chinese listeners.

## 2. Computer model

### 2.1. Corpus

The computer model was built using isolated vowel-consonant-vowel (VCV) tokens derived from material collected for the Interspeech 2008 Consonant Challenge (Cooke and Scharenborg 2008), which includes all 24 English consonants in vowel contexts derived from all nine combinations of /æ/ /iː/ /uː/ as first and second vowel, produced by 12 female and 16 male speakers. A similar Mandarin Chinese VCV corpus including all 24 Chinese consonants (/$p^h$ p $t^h$ t $k^h$ k $ts^h$ ts $\widetilde{tʃ^h}$ $\widetilde{tʃ}$ $\widetilde{tɕ^h}$ $\widetilde{tɕ}$ f s ʃ ɕ x m n ŋ l ɻ̈ j w/) in the same set of vowel contexts was collected from 12 female and 17 male speakers, all native Chinese students studying at the University of Sheffield. Post-processing involved high-pass filtering to attenuate low frequency energy below 50 Hz from the tokens, followed by endpointing to remove silence. Tokens were downsampled to 25 kHz and normalised to have the same root mean square energy level. The corpus was screened manually to identify and remove tokens which were incorrectly

produced or contained noise from, for example, key tapping. Following screening, a total of 3299 English and 3331 Chinese tokens were available for model training and testing.

## 2.2. Hidden Markov models

Separate models were built for English and Chinese sounds using speech material from the two corpora. Unlike Thomson et al. (2009), who used discriminant analysis to build vowel recognition models, we built our models by using the standard Hidden Markov modelling (HMM) techniques and acoustic representations (Mel-Frequency Cepstral Coefficients) employed in automatic speech recognition. Models were constructed and trained using HTK, the Hidden Markov Model Toolkit (Young et al. 2006).The baseline recognition models were trained for each consonant and vowel in each language, resulting in two recognition systems, one for each language. Hidden Markov models essentially compute an optimal assignment or clustering of acoustic data to model state sequences. Here, each sound (consonant or vowel) was represented by three states, and each individual state learns a statistical distribution of acoustic information allocated to that state. The allocation of training data to states is determined by within-state similarity rather than uniformly sampling the interval during which the consonant or vowel is present. Thus, for example, while the model for a plosive is likely to contain a sequence of states representing formant transition, closure and burst, these segments are not manually-identified but emerge automatically from similarity criteria during the training phase. Trained HMMs can be used to categorise a new instance of a sound, whether or not that instance belongs to the language for which the models were trained. The categorisation result is obtained by choosing the HMM with the highest likelihood of generating the sound to be classified. The rightmost column of Table 1 provides classification results for English VCVs using HMMs trained on Chinese VCVs. We discuss these results in section 4.

## 2.3. Cross-language model distance

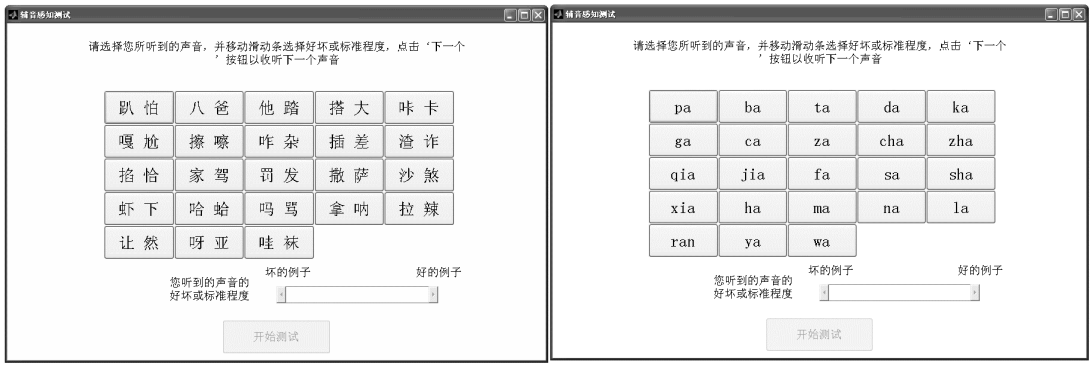
Cross-language sound similarity measures can be generated by taking all possible pairings of HMMs trained on English or Chinese sounds. Distances are not based on single exemplars of English and Chinese sounds but on the acoustic distributions learnt by the HMMs. Consequently, the distance measures we use involve the overlap between pairs of probability distributions rather than conventional pointwise distances such as the Euclidean or cityblock metrics. A number of distance measures employed in pattern recognition were used in current study to calculate all pairwise distances between the consonant HMMs of the two languages. These measures including Bhattacharrya distance, which is a

measurement of the similarity of two probability distributions; Kullback-Leibler divergence, also called Relative Entropy in information theory, which is also a distance measurement between two probability distributions. In addition, Kullback-Leibler divergence with Monte-Carlo sampling was also used, which differs from the standard Kullback-Leibler divergence in using randomly-generated sample points of two probability distributions rather than probability density functions themselves (Mak and Barnard 1996; Sooful and Botha 2002; Cover and Thomas 2006; Hershey and Olsen 2007).

## 3. Cross-language mapping experiments

Human perception data were collected to test and refine the computer model. Native Chinese listeners categorised English consonants in terms of their native sounds. To afford comparison with model distance measures, listeners also provided a goodness rating. Chinese is a well known language using the logographic writing system. Its characters are not directly linked to their pronunciations. However, Chinese has a Romanisation system called Pinyin, which uses Roman letters to mark the sounds and to help the learner of Chinese to remember the pronunciation of the Chinese characters. Most Chinese children learn Pinyin in primary school at the same time as they learn Chinese characters. To examine whether the differing orthographic forms of Chinese characters or Pinyin symbols influence the categorisation or goodness ratings, one group of participants responded using Chinese characters while the other group used Pinyin (see Figure 1).

Figure 1: Screen shots of the experiment interface (left: character group; right: Pinyin group).



### 3.1. Stimuli

A subset of English and Chinese VCV tokens used to train the computer model were used as stimuli in the cross-language mapping experiments. 552 English tokens from 16 male speakers, including all English consonants except /ŋ/ (excluded due to phonotactic considerations) were used in the experiments. The initial vowel was one of /æ, iː, uː/ while the final vowel was /æ/. All tokens has end stress. 46 Chinese tokens selected from those produced by 12 male speakers were mixed into the English tokens. A further 23 Chinese tokens were used in a prior keyboard training session. The Chinese tokens used the same vowel contexts as the English. The average length of the final vowel /æ/ in Chinese VCVs (168 ms) is significantly shorter than the VCV-final English /æ/ (233 ms), as also found Guion et al. (2000). To prevent the use of final vowel length as a cue in the perception experiments, we also normalised the length of the final vowel of all the English and Chinese tokens to 150 ms, with the final 20 ms linearly ramped down to zero amplitude.

### 3.2. Participants

Thirty native Chinese speakers, 10 females and 20 males, were divided into two listener groups of 15 participants. One group used Chinese character symbols to record their perceptions, while the other group used Chinese Pinyin symbols. All were second year students (mean age = 20.6) at the Xi'an Technological University, China, studying non-linguistic courses, and none had lived outside China. Children start to learn English in China when they go to middle school and sometimes earlier. It is very difficult to find Chinese young adults who have never had English lessons. On average, our participants had begun to learn English at 12.6 years. Most of their exposure to English was in the classroom with Chinese teachers, and sometimes from other sources such as music and films. Most of the participants were from north-west China, which is the same dialect area as Mandarin Chinese. Pure-tone hearing tests at frequencies from 250 to 8000 Hz were carried out using a software audiometer. One participant from the character group was excluded due to problems in both ears at high frequencies. All participants received a small payment for taking part.

### 3.3. Procedure

The experiments were carried out in a quiet meeting room in Xi'an Technological University. Participants were tested individually and heard stimuli via Sennheiser HD650 headphones and an M-Audio Mobilepre external sound card. Stimulus presentation and response collection was controlled by a computer program. Participants used the mouse to select their response category buttons on a virtual

keyboard, with Chinese characters or Pinyin graphemes as the symbol of the VCV tokens on each button (Figure 1). Participants were asked to first classify the token they heard as an instance of one of the Chinese consonant categories, then to move a slider bar under the keyboard to give a goodness rating of the sound (relative to the Chinese category they selected). The goodness rating used a 0-100 scale (0=bad, 100=good), although participants were only aware of the continuous sliding scale. Both a keyboard training session and a practice test were carried out before the formal test. Participants were asked to click each button to hear the corresponding Chinese token (1 for each button) to familiarise themselves with the Chinese consonants and their positions on the keyboard. The practice test contained two examples of each English consonant (46 tokens in all). In the formal test, 46 Chinese tokens (2 for each consonant) and 506 English tokens (22 for each consonant) were mixed and presented to the participants in 3 sessions (184 tokens randomly distributed for each session). Participants were asked to take a short break before they continued to the next session.

## 4. Results

### 4.1. General description

The cross-language mapping results are listed separately for the Chinese character and Pinyin groups in Table 1. The cross-language computer model recognition results are also shown. Some English consonants, especially the plosives, were assimilated to certain Chinese categories with a very high frequency, and the goodness ratings for these English sounds were also very high, which suggests that the distances between these sound-pairs are relatively small. Other sounds, especially those English sounds which are known to lack Chinese counterparts, produced more confusions, and the goodness ratings were also lower. Inspired by Guion et al. (2000), a 'fit index' was calculated by multiplying the percentage and the goodness rating. The mean fit indices for the Chinese consonants used as an experiment control (section 3.1) were 69.8 (sd: 16.3) for the character group and 71.4 (sd: 16.7) for the Pinyin group. Based on the same standard deviation criterion as Guion et al. (2000), English consonants were classified into 'good', 'fair' or 'poor' instances of Chinese categories, as depicted in grey tones in the table. Four pairs of English consonants – /f v/, /θ s/, /ð z/ and /ʃ ʒ/ – where both sounds in each pair were assimilated to the same one or more Chinese categories, are expected to be difficult for Chinese listeners to discriminate.

4.2. Listener – Model comparison

Table 1 also shows that the model's cross-language recognition results are consistent with listeners' categorisation results. While listeners made fewer confusions, if we focus on the first few major confusions, the model made almost the same confusions as listeners. If the confusion ranking is considered, the biggest differences lie in the English consonants /ʧ ʤ ʒ/. Interestingly, in these cases the best ranked confusion for listeners was second ranked by the model and vice versa. Also, the percentage differences between the first two confusions in the model were smaller than for listeners. This suggests that listeners are able to exploit information which is poorly-represented or missing in the model's acoustic representation or topology.

As mentioned in section 2.3, the model permits an estimation of between-language consonant distance. Table 2 lists the correlations between model distances and listener goodness ratings. The two-way Kullback-Leibler divergence with Monte-Carlo sampling (KL_MC2) is the distance measurement with the most significant and highest correlation with the goodness rating [$r = -0.66$, $p < .001$]. The group presented with Chinese characters showed higher correlations with the model than the Pinyin group. Since model distances are purely acoustic, this finding points to orthographic influences from the Pinyin characters.

Table 1: Results of the cross-language mapping experiments and the cross-language computer model recognition. For each listener group, mean categorisation percentages and goodness ratings are provided as well as a fit index described in the text. Categorisation percentages from the model are also shown. Only the principal confusions (≥5%) are shown in the table. Rows highlighted in dark are 'good' exemplars, those in light grey are 'fair', while the rest are 'poor'.

| English | Character | | | Pinyin | | | Model |
|---|---|---|---|---|---|---|---|
| | % | good | fit | % | good | fit | |
| p | pʰ(99%) | 80 | 79.2 | pʰ(98%) | 82 | 80.36 | pʰ(61%) p(20%) tʰ(9%) t(6%) |
| b | p(98%) | 77 | 75.46 | p (97%) | 78 | 75.66 | p (56%) t(23%) w(8%) |
| t | tʰ(86%) | 75 | 64.5 | tʰ(82%) | 75 | 61.5 | tʰ(23%) ʧʰ(20%) ʨʰ(17%) |
| | tsʰ(9%) | 72 | 6.48 | tsʰ(13%) | 76 | 9.88 | tsʰ(10%) ʧ(10%) ts(7%) t(6%) |
| d | t(95%) | 75 | 71.25 | t(98%) | 76 | 74.48 | t(68%) k(21%) ʧ(8%) |
| k | kʰ(95%) | 78 | 74.1 | kʰ(95%) | 78 | 74.1 | kʰ(52%) tʰ(16%) k(13%) ʨʰ(6%) |
| g | k(97%) | 79 | 76.63 | k(98%) | 77 | 75.46 | k(76%) t(9%) n(5%) |
| ʧ | ʧʰ(60%) | 71 | 42.6 | ʧʰ(85%) | 72 | 61.2 | ʨʰ(55%) ʧʰ(23%) ʧ(15%) |
| | ʨʰ(14%) | 70 | 9.8 | ʧ(7%) | 64 | 4.48 | |
| | tsʰ(13%) | 69 | 8.97 | ʨ(5%) | 52 | 2.6 | |
| | ʧ(7%) | 71 | 4.79 | | | | |
| ʤ | ʧ(63%) | 75 | 47.25 | ʧ(69%) | 71 | 48.99 | ʨ(38%) ʧ(32%) ʨʰ(11%) |
| | ʨ(25%) | 70 | 17.5 | ʨ(21%) | 66 | 13.86 | ɻ(10%) |
| f | f(97%) | 77 | 74.69 | f(99%) | 80 | 79.2 | f(61%) ʃ(14%) s(13%) x(9%) |
| v | w(74%) | 64 | 47.36 | w(75%) | 68 | 51 | ɻ(27%) f(19%) ʃ(10%) |
| | f(19%) | 60 | 11.4 | f(20%) | 46 | 9.2 | x(12%) s(8%) p(5%) |
| θ | s(50%) | 68 | 34 | f(56%) | 70 | 39.2 | f(35%) s(32%) ʃ(13%) |
| | f(46%) | 74 | 34.03 | s(41%) | 63 | 24.6 | x(11%) |
| ð | w(30%) | 60 | 18 | ts(45%) | 58 | 26.1 | ts(18%) ɻ(14%) ʃ(12%) |
| | ts(20%) | 43 | 8.6 | w(34%) | 56 | 19.04 | f(11%) x(8%) t(8%) s(7%) |
| | ɻ(20%) | 44 | 8.8 | f(11%) | 39 | 4.29 | |
| | f(13%) | 53 | 6.89 | | | | |
| | s(8%) | 45 | 3.6 | | | | |
| | ʧ(6%) | 45 | 2.7 | | | | |
| s | s(96%) | 78 | 74.88 | s(95%) | 81 | 76.95 | s(55%) ʃ(28%) ɕ(10%) |
| z | ɻ(42%) | 45 | 18.9 | ts(74%) | 71 | 52.54 | ɻ(34%) s(22%) ʃ(12%) |
| | ts(33%) | 42 | 13.86 | ɻ(13%) | 44 | 5.72 | ɕ(12%) ts(6%) j(5%) |
| | s(18%) | 45 | 8.1 | s(6%) | 57 | 3.42 | |
| ʃ | ʃ(83%) | 80 | 66.4 | ʃ(91%) | 78 | 70.98 | ʃ(69%) ɕ(22%) ʨʰ(7%) |
| | ɕ(14%) | 66 | 9.24 | ɕ(5%) | 67 | 3.35 | |
| ʒ | ɻ(68%) | 51 | 34.68 | ɻ(78%) | 56 | 43.68 | ʃ(39%) ɻ(25%) ɕ(15%) |
| | ʃ(14%) | 58 | 8.12 | ʃ(10%) | 63 | 6.3 | ʨʰ(6%) |
| | j(9%) | 58 | 5.22 | | | | |
| h | x(98%) | 78 | 76.44 | x(98%) | 82 | 80.36 | x(79%) |
| m | m(96%) | 78 | 74.88 | m(98%) | 78 | 76.44 | m(88%) l(5%) |
| n | n(88%) | 74 | 65.12 | n(96%) | 75 | 72 | n(56%) m(27%) l(9%) |
| | l(11%) | 75 | 8.25 | | | | |
| l | l(89%) | 79 | 70.31 | l(97%) | 82 | 79.54 | l(69%) ɻ(12%) n(6%) w(6%) |
| | n(10%) | 71 | 7.1 | | | | |
| r | ɻ(83%) | 49 | 40.67 | ɻ(93%) | 55 | 51.15 | ɻ(61%) w(15%) l(8%) j(5%) |
| | w(11%) | 46 | 5.06 | | | | |
| j | j(99%) | 80 | 79.2 | j(98%) | 80 | 78.4 | j(67%) ɻ(12%) w(9%) |
| w | w(96%) | 78 | 74.88 | w(96%) | 77 | 73.92 | w(80%) x(6%) |

4.3. Character group – Pinyin group comparision

Table 1 suggests that most of the classifications by the two groups are similar. However, differences occurred for the English consonants /ʧ θ ð z ʃ ʒ n l r/**.** For instance, English /ʧ/ was heard as Chinese /ʧʰ/ on 85% of occasions by the Pinyin group compared to 60% for the character group, while figures for English /z/ heard as Chinese /ts/ are 74% and 33% respectively. Chinese characters form a logographic system, whose written form is not directly linked to the pronunciation, while Pinyin is a Romanised alphabetic system, which is used to represent the pronunciation of Chinese characters. For the English sounds /ʧ z ʃ r/**,** this "pronunciation-recall" aspect of Pinyin may have had an influence. In Pinyin, the Chinese consonants /ʧʰ ts ʃ ɻ/ are written as "ch", "z", "sh" and "r", which have the same written forms as the English consonants /ʧ z ʃ r/. As all the participants were university students and knew some English (although none were fluent), and they knew the pronunciation of those Pinyin symbols in English. Although participants were not told the language of sounds they would hear, since English is the only foreign language they knew, it was very likely that they would connect the sounds they heard to English when making their choice and supplying a rating. For sounds such as /ð/ and /ʒ/, Pinyin may have had an orthographic influence on category decisions. Another interesting case concerns the /n/-/l/ confusions made by the Character group but not the Pinyin group. A closer look at the participants' individual data shows that most of the /n, l/ confusions came from 3 participants whose dialect belongs to the south west Mandarin region where this confusion is common.

Table 2: Correlations between goodness ratings and distance measurements for the Character and Pinyin groups. Both raw and normalised (z-score) goodness ratings are provided. Bha: Bhattacharrya distance; KL2: 2-way Kullback-Leibler divergence; KL_MC: KL divergence + Monte-Carlo sampling, KL_MC2: 2-way KL divergence + Monte-Carlo sampling.

| Raw goodness rating | | Distance measurements | Normalised goodness rating | |
|---|---|---|---|---|
| Character | Pinyin | | Character | Pinyin |
| –0.29 n.s. | –0.20 n.s. | Bha | –0.31 * | –0.24 n.s. |
| –0.34 * | –0.12 n.s. | KL2 | –0.36 * | –0.16 n.s. |
| –0.64 *** | –0.49 ** | KL_MC | –0.64 *** | –0.48 ** |
| –0.66 *** | –0.54 *** | KL_MC2 | –0.67 *** | –0.53 ** |

5. Discussion

The listener–model comparison shows clear similarities in sound categorisation and significant (negative) correlation between listeners' goodness ratings and

model sound-pair distances, suggesting that acoustic clustering techniques used in automatic speech recognition may be valuable in cross-language studies. For example, it would be possible to predict which L2 sounds would be problematic for any given set of L1 listeners by training on speech material from the two languages. The modelling approach also has the potential to provide a more fine-grained classification of sound similarity than conventional qualitative models. However, listener-model differences in classification rates even for the categories which have a clear cross-language assimilation point to imperfections in the current computer model's ability to accurately represent those aspects of the acoustic signal which listeners have access to in decision making. The failings of the hidden Markov modelling framework are widely-acknowledged (e.g. HMMs are poor at duration modelling). Further, the acoustic representation used here is well-matched to HMMs but is almost certainly not that used by human listeners. Another limitation of the computer model is that it is trained on VCV tokens rather than natural speech. By acquiring speech using more natural material, listeners are exposed to more variety, for instance, in context, speech rate and accent. This limitation will be overcome in future work by training with more natural material. Possible orthographic influences revealed by Character-Pinyin differences highlights a methodological concern: as more young people in China start to learn English very early in their lives, use of Pinyin symbols in cross-language mapping experiments may become less reliable. However, Chinese characters are not free from problems, since they can have different pronunciations in different dialect regions, or even in the same dialect region by listeners with different backgrounds, as evidenced by the /n, l/ confusions which occurred for the Character group.

## REFERENCES

Best, C. 1995. A Direct Realist View of Cross-Language Speech Perception. In: W, Strange. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York, Timonium, MD. 171–204.

Cooke, M. P. and Scharenborg, O. 2008. The Interspeech 2008 Consonant Challenge. *Proc. Interspeech*, *Brisbane, Australia*.

Cover, T. M. and Thomas, J. A. 2006. *Elements of Information Theory*, Second Edition. John Wiley and Sons.

Flege, J. 1995. Second-Language Speech Learning: Theory, Findings and Problems. In: W, Strange. (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. York, Timonium, MD. 233–273.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., Pruitt, J. C. 2000. An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *J. Acoust. Soc. Am.* 107(5). 2711–2723.

Hershey, J. R., Olsen, P. A. 2007. Approximating the Kullback-Leibler divergence between Gaussian mixture models. *Proceedings of ICASSP 2007, Honolulu, USA*. 317–320

Lengeris, A., Hazan, V. 2007. Cross-language perceptual assimilation and discrimination of southern British English vowels by Greek and Japanese learners of English. *Proceedings of the 16th ICPhS, Saarbrücken*. 1641–1644.

Mak, B. and Barnard, E. 1996. Phone clustering using the Bhattacharyya distance. *Proc. ICSLP'96, Philadelphia, USA*. 2005–2008.

Sooful, J. and Botha, E. 2002. Comparison of acoustic distance measure for automatic cross-language phoneme mapping. *Proc. ICSLP'2002, Denver, USA*. 521–524.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S., Nishi, K., and Jenkins, J. 1998. Perceptual Assimilation of American English Vowels by Japanese Listeners. *J. Phonetics* 26. 311–344.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S., and Nishi, K. 2001. Effects of consonantal context on perceptual assimilation of American English vowels by Japanese listeners. *J. Acoust. Soc. Am.* 109(4). 1691–1704.

Thomson, R. I., Nearey, T. M., Derwing, T. M. 2009. A modified statistical pattern recognition approach to measuring the crosslinguistic similarity of Mandarin and English vowels. *J. Acoust. Soc. Am.* 126(3). 1447–1460.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. *The HTK book (for HTK version 3.4)*. Cambridge: Cambridge University Engineering Department.