音声研究　第 19 巻第 1 号
2015（平成 27）年 4 月
1–12 頁

Journal of the Phonetic Society
of Japan, Vol. 19 No. 1
April 2015, pp. 1–12

特集論文

Exploring the Representation of Second Language Sounds in First Language Space

Maria Luisa Garcia Lecumberri*, Jian Gong* and Martin Cooke*

和文タイトル

和文著者名*

SUMMARY: The extent to which language learners hear non-native sounds in terms of native categories depends in part on acoustic and auditory similarities between the two sets of sounds. One unresolved issue is the choice of parameter space in which similarity should be measured. The current paper demonstrates the application of an unsupervised, corpus-based, data-driven mapping technique which permits the use of rich, high-dimensional data representations, obviating the need for prior commitment to specific low-order speech parameters such as formant frequencies. The approach, known as generative topographic mapping, preserves the structure of the high-dimensional space while mapping to a lower-dimensional space. We show how this low-dimensional latent space can be used for tasks such as visualising the location of L2 consonants in an existing L1 space and measuring the effect of L2 exposure on the representation of both L2 and L1 consonants by comparison with data from a behavioural study in which Chinese listeners underwent an intensive training regime on Spanish consonants.

Key words: L1, L2, consonants, unsupervised mapping, latent space, assimilation, attrition, model

1. Introduction

Listeners learning a second language (L2) are usually exposed to substantial quantities of foreign language speech from instructors, educational materials, television and radio as well as conversational partners. Since the origin of learners' difficulties with L2 sounds is, to some extent, grounded in the foreign and native language speech signals themselves, one approach to understanding the nature of L2 perception is to undertake corpus-based studies. Further, variability due to factors such as talker differences and articulatory context demands the use of large samples of speech material. The chief difficulty in large-scale corpus-based studies of L2 speech perception is to find an appropriate transformation from the high-dimensional space of the signals themselves to a representation in which sounds from the first and second languages can be compared.

Corpus-based studies linked to powerful statistical learning and data modelling algorithms complement theoretical models of L2 sound acquisition (e.g. Best 1995, Flege 1995, Kuhl 1993) by generating quantitative predictions of the degree to which L2 sounds will be difficult to acquire.

The current article applies an unsupervised statistical learning algorithm — Generative Topographic Mapping (GTM; Bishop et al. 1998) — to the problem of visualising and quantifying the similarity between L1 and L2 sounds. The GTM learns a low-dimensional 'latent' space directly from speech data. Once the latent space for the L1 sounds is constructed it is possible to answer such questions as: (i) what is the spatial structure of L2 sounds in the L1 space? (ii) how does the relationship between L1 and L2 sounds change as a response to increasing exposure to L2 sounds? and (iii) how do the locations of L1 sounds themselves change as a result of L2 sound exposure? These questions relate directly to three main issues of L2 sound acquisition, namely L1/L2 assimilation, L2 categorisation and L1 attrition. In addition to visualisation, the latent space of the GTM can be used to generate quantitative predictions by, for instance, computing distances between centroids representing the locations of L1 and L2 sounds in the network.

Section 2 reviews earlier corpus-based approaches in both first and second language acquisition. Sec-

* Language and Speech Lab, University of the Basque Country, Spain（バスク大学）

tion 3 describes the Generative Topographic Mapping technique employed to construct the latent space of L1 sounds. The remainder of the article describes the outcome of an experimental investigation into the representation of Spanish sounds in a latent space of Chinese sounds.

2. Corpus-based studies in L1 and L2 sound acquisition

2.1 Modelling L1 sound acquisition

Native perception studies have employed corpus-based computational and statistical modelling techniques, particularly in investigations of the native language perceptual magnet effect and native category formation (Kuhl 1993). For example, Iverson and Kuhl (1996) used multidimensional scaling to process listeners' identification and goodness rating data, and demonstrated a clear perceptual magnet effect, that is, a shrinking perceptual space near category prototypes and a stretching near category boundaries. A study by de Boer and Kuhl (2003) extracted vowel formant information from the words produced by care-givers and used them as input to a Gaussian Mixture Model (GMM) to simulate infant acquisition of native phonetic categories from the ambient language environment. Vallabha et al. (2007) used F1, F2 and duration information derived from mothers' naturally-produced speech to create Gaussian models for different vowels, and used these models to generate a large amount of training data tokens for their online mixture estimation model of infants' native vowel category formation. The learning process used in this model can acquire vowel categories on a token-by-token base, a method more similar to the everyday situation facing infants when acquiring categories utterance-by-utterance from care-givers.

Many of the modelling studies cited above are based on the parametric GMM framework, but other studies have attempted to explore the underlying mechanisms of categorical learning through artificial neural-network techniques. For example, Guenther and Gjaja (1996) adopted the Self-Organizing Map (SOM; Kohonen 1988) to simulate vowel space formation and other aspects of the native perceptual magnet effect, based on formant information from vowel and consonant categories. Vallabha et al. (2007) also proposed an alternative neural-network version of their model. Salminen et al. (2009) used a self-organized network which employed the Hebbian learning rule (Hebb 1949) to model the warping of perceptual space associated with categorical perception. The input data of Salminen et al. (2009) were synthesized vowel sounds processed through a model of the auditory periphery, unlike the self-organising model of vowel category acquisition described in Miyazawa et al. (2010) which used mel-frequency cepstral coefficient (MFCC) vectors derived from natural continuous speech.

2.2 Modelling L2 sound acquisition

Corpus-based computational and statistical techniques have also been employed in non-native speech perception studies. For example, Strange et al. (2004) adopted a linear discriminant analysis approach (LDA; Klecka 1980) to measure acoustic similarities between American English and North German vowels, using the formant values F1, F1 and F3 extracted from speech tokens in isolated citation forms and in sentence context. Morrison (2006) extended LDA to a canonical discriminant function analysis approach (CDFA; Johnson 1998, Tatsuoka 1970) to model how Spanish and English listeners perceive vowels in each other's language. One important merit of the CDFA technique — shared with the GTM approach we describe below — is the ability to map the characteristics of the data from a high dimensional acoustic space to a lower dimensional representation, thereby providing the possibility of visualisation. Specifically, Morrison (2006) mapped from a 5-dimensional data space (F1, F2, Δ F1, Δ F2, duration) to a 2-dimensional visualisation space.

A similar statistical pattern recognition technique was applied by Thomson et al. (2009) in their cross-language modelling study aimed at measuring the similarity between Mandarin and English vowels using formant information. Thomson et al. (2009) built recognition models for both Mandarin and English vowels based on LDA, placing them in competition with each other during the recognition phase. The distribution of recognition percentages between the competing models was used to construct an index of the similarities between the vowel categories that those models resembled. By taking the a posteriori probability[1)] obtained from the LDA as a simulation of the goodness of fit scores, the model of Thomson et al. (2009) can simulate behavioural responses in cross-language assimilation tasks.

Computational approaches also have been used in studies of the development of L2 perception. Escudero et al. (2007) used machine learning algorithms (k-nearest neighbour, naive Bayes classifier) and computational linguistic models (the gradual learning algorithm) to simulate and visualise the evolution of learn-

ers' L2 vowel spaces, based on learners' category perceptual data for synthetic stimuli. Hidden Markov modelling techniques were adopted by Gong et al. (2011a) in a modelling study investigating the effect of very limited amounts of exposure on learning L2 consonants, using MFCC-based acoustic feature representations derived from natural speech data.

3. Generative topographic mapping

Generative topographic mapping (Bishop et al. 1998) creates probability density models in a low-dimensional latent space and maps them to the high-dimensional observational data space via a smooth mapping function (e.g. a radial basis functions[2] network; Broomhead and Lowe 1988). The GTM approach is motivated by similar concerns as those that led to the SOM (Kohonen 1988), namely unsupervised learning from large datasets resulting in a projection to a lower-dimensionality space for ease of visualisation. However, adaptation in the SOM is driven by heuristics rather than statistical learning considerations, and lacks, for example, a principled basis for the choice of learning rate and neighbourhood functions[3]. The parameters of the density models in the GTM latent space can be estimated and optimised by measuring the parameters of the density models in the high-dimensional data space through the expectation-maximisation (EM) algorithm[4] (Dempster et al. 1977). In this way the topographic characteristics of the data in the high-dimensional space can be learned and preserved in the lower dimensional latent space, where their interpretation is more tractable.

It is the ability to visualise the low-dimensionality 'intrinsic' features that characterise consonants in a given language that motivates the use of the GTM as an analysis tool in second language studies. It is also feasible to examine the evolution of the space with increasing amounts of L2 input. Further, the GTM procedure is wholly data-driven (i.e. unsupervised) and hence requires no category labels for the observations which form the input to the algorithm. Portions of the latent space receive category labels only after learning is complete. This feature permits the examination of the role of pure acoustic similarity between L1 and L2 sounds without top-down influences such as phoneme categories or orthography. A further benefit of the unsupervised approach is in allowing the use of large unlabelled speech corpora, although labels are required for a fraction of the corpus in order to interpret the space following learning, as described below.

The application of the GTM technique to the representation of L2 consonants in L1 space proceeds as follows. First, the latent space for the L1 consonants is constructed using the GTM algorithm from a large collection of speech parameter vectors. Second, the topography of the latent space is visualised by examining the response probabilities at each location in the space for a subset of labelled L1 consonants. The result is a map (usually restricted to 2 or 3 dimensions) showing both the location and extent of each consonant in the latent space. Once the L1 space has been mapped, the locations of L2 consonants can be plotted using their response probabilities, revealing the purely acoustic relationship between L1 and L2 sounds. The GTM construction procedure can be repeated using the original L1 consonant set along with increasing amounts of L2 input to simulate the effect of increasing exposure, and the locations of both L1 and L2 consonants within the new latent space are examined.

4. Application of the GTM to localising Spanish sounds in Chinese consonant space

This section illustrates the GTM procedure for the case of situating (i.e. acquiring by exposure) Spanish consonants in the latent space of Chinese consonants. This language pairing was chosen due to the availability of speech corpora of comparable size and acoustic context for both Spanish and Chinese, as well as the existence of behavioural results on Chinese assimilation and categorisation of Spanish consonants (Gong 2013). Further, Chinese and Spanish are the two most-spoken languages by native talkers and exhibit large differences in their phoneme inventories and realisations. For example, the plosives are contrasted by aspiration in Chinese and voicing in Spanish; Chinese has a large affricate inventory compared to the single affricate phoneme in Spanish; and the well-known liquid difficulties experienced by Chinese listeners with languages such as English are compounded by the presence of two 'r' sounds, the tap /ɾ/ and the trill /r/, both different realisations from the Chinese apical post-alveolar approximant /ɻ/.

4.1 Corpora

Naturally-produced VCV tokens in both Mandarin Chinese and Spanish were used to construct and evaluate the mapping created using GTM. Chinese VCV tokens were drawn from the corpus described in Gong et al. (2011b). This corpus is made up of exemplars of the 24 Chinese consonants shown in the first column

Table 1 Chinese and Spanish consonant inventories.

	Chinese	Spanish
Plosive	pʰ p tʰ t kʰ k	p b t d k ɡ
Fricative	f s ʂ ɕ x	f θ s x
Affricate	tsʰ ts tʂʰ tʂ tɕʰ tɕ	tʃ
Nasal	m n ŋ	m n ɲ
Liquid	ɻ l	l ɾ r
Approximant	j w	j

of Table 1 produced in 9 intervocalic contexts derived from all combinations of the vowels /a, i, u/ in initial and final vowel position. The Chinese corpus contains a total of 3,331 VCV tokens from 17 male talkers after removal of noisy or mis-produced tokens.

Spanish VCV tokens were taken from the corpus collected for the behavioural studies described in Gong (2013). This corpus contains the 18 Spanish consonants shown in the second column of Table 1 in the same 9 vowel contexts as used in the Chinese VCV corpus. A total of 3,240 tokens from 16 male speakers were recorded for the Spanish corpus. Of these, 2,880 represent training tokens (10 exemplars of each of the 18 consonants in each of 16 sessions; see section 4. 3. 2) and a further 360 were used for testing (20 exemplars of each consonant).

Tokens from the two corpora were subject to identical post-processing: down-sampling to 25 kHz, high-pass filtering to remove energy below 50 Hz, and normalisation of RMS energy.

4. 2 Feature extraction

Only the consonant part of each VCV token was used for modelling. The boundaries of the consonant portion of each VCV were identified by HMM-based forced alignment[5] using HTK (Young et al. 2006). Every 10 ms a 39-dimensional feature vector was formed from the first 12 MFCCs plus overall energy, together with their first and second time derivatives. To accommodate different consonant durations within the constraints of a fixed size parameter vector, and to capture the temporal dynamics of each consonant, MFCC parameters at frames located closest to 25, 50 and 75% of the consonant interval were concatenated to form a single 117-component feature vector per consonant token. These observational data vectors form the input to the GTM.

4. 3 GTM training and visualisation

4. 3. 1 Learning of the L1 latent space

Since initial experiments suggested that a 2-dimensional latent space produced too large an overlap between consonants, a 3-dimensional GTM was constructed. Following pilot studies the latent space was represented by 225 probability density model centres or 'nodes' distributed randomly in the space defined by the cube with vertices at [±1,±1,±1] centred on the origin. The radial basis function network which served as the projection function was set to have 72 nodes in order to obtain a smooth mapping from latent space to data space. The Netlab GTM toolbox (Nabney 2004) was used for model construction.

The GTM was trained on the entire Chinese corpus (i.e. 3,331×117-D observation vectors) to simulate a Chinese consonant space prior to Spanish exposure. Fifty iterations of EM in the unsupervised GTM learning algorithm were sufficient to obtain convergence.

4. 3. 2 Simulating the effect of L2 exposure

While small quantities of L2 data seem unlikely to cause significant disturbance to the locations of L1 consonants, increasing the quantity of L2 tokens might be expected to result in changes to the latent space constructed by the GTM. Putative changes might involve a greater dispersion of L2 categories, or increased sensitivity (in some latent space dimensions) to cues not used in certain L1 distinctions. Since the L1 and L2 sound systems coexist in the same phonological space (Flege 1995), through increased exposure to the L2 sound system, the perceptual space may change by creating new categories for some L2 sounds, particularly if the differ noticeably from L1 categories; when that happens, the existing L1 category may shift in the perceptual space to better differentiate itself from a newly created L2 category (dissimilation; Flege 2002). Alternatively, an L2 sound may get assimilated to an existing L1 category, in which case the existing L1 category may be modified to accommodate the L2 one (assimilation; Flege 2002). For example, in the acquisition of Spanish voiced plosives, Chinese learners may create new categories for these sounds with or without shifting their native unaspirated plosives to increase their distance. Learners could also assimilate Spanish voiced plosives to Chinese unaspirated plosives, in which case the Chinese categories might change through the influence of the Spanish sounds they have incorporated.

In the behavioural study of Gong (2013), listeners were trained sequentially on 16 blocks of 180 Spanish VCV tokens made up of 10 exemplars of each consonant. To simulate the effect of increasing exposure in the GTM, an identical subdivision into blocks was employed. For each of the 16 stages of learning, all the Spanish blocks up to that stage were added to the Chinese training corpus. A further 10 iterations of EM were

carried out to modify the GTM parameters using the gradually-expanding training corpus. The GTM parameters obtained in the previous stage were used as input parameters for the next stage of EM-based parameter re-estimation. The pure Chinese GTM served as the input for the first Spanish learning stage. In this way, the initial Chinese GTM gradually gained both an increased amount and more varied Spanish exposure while keeping its Chinese exposure unchanged, simulating a listener's non-native sound learning.

4.3.3 Visualisation of the latent space

In order to identify the locations of consonants in the GTM latent space, the response probability at each GTM node is computed for each labelled feature vector in the test set. Rather than choosing the most likely node, the 3-D location of the feature vector is obtained by multiplying the response probability matrix (i.e. the matrix of response probabilities defined across all GTM nodes) by the coordinate matrix of the nodes. In this way, each labelled feature vector is mapped into a 3-D coordinate in the latent space. This can be thought of as representing the centre of activity in the map for that token.

Ellipsoids are used to visualise the variability of each individual consonant's response distribution across all tokens of the same consonant. Ellipsoid centres are placed at the mean 3-D location of the responses, their radii represent one standard deviation in each of the three dimensions, and their 3-D orientation is computed from the eigenvector matrix of the 3-D coordinates matrix of the mean responses.

5. Results

5.1 Consonant distribution in latent space

Figure 1 shows the spatial distribution of individual Spanish and Chinese consonants in the Chinese GTM before, during and after training on Spanish consonants. In this figure colour distinguishes different manners of articulation; ellipses representing individual consonants are not labelled to avoid clutter (below we show projections for individual consonants from 3-D space to pairs of dimensions).

Considering first the distribution of Chinese consonants prior to exposure to Spanish (upper left panel), a clear division exists between the sonorant consonants (nasals and approximants) and the obstruents (plosives, fricatives, affricates). In fact, for Chinese, the sonorant-obstruent distinction is also a voiced-voiceless contrast, since Chinese does not contain voiced plosives, fricatives or affricates. Thus, the GTM trained solely on Chinese consonants seems to be sensitive to either or both of the sonorant-obstruent and voiced-voiceless distinctions.

The spatial distribution of Spanish consonants in the Chinese GTM prior to Spanish exposure is depicted in the upper right panel of Figure 1. The consonants are more tightly clustered than seen for the Chinese consonants, although some separation of manner classes is evident, especially for the affricates, fricatives and plosives. Note that in the Chinese GTM the liquids are not visible as they are subsumed within the ellipsoid representing approximants.

One measure of clustering is the mean inter-consonant distance: under this metric, the Spanish consonants are 76% less well-separated than the Chinese consonants. Comparing the locations of the Chinese and Spanish consonants in Figure 1, the largest difference is seen for the plosives and fricatives. Additionally, Chinese has a large area for approximants which encompasses the liquids. Spanish has a specific extensive area for the liquids. We provide a more detailed comparison of individual consonants in the two languages below.

As exposure progresses (session 4; right column, centre panel) some movement in the location of Spanish consonants is evident, and at the termination of exposure (session 16; lower right) the extent of their movement is clear. In fact, the mean Spanish consonant separation is now 6% greater than that of the Chinese consonants following exposure. In general, the manner classes are better separated, with the liquid grouping emerging from the plosive cluster. A similar picture is seen for some of the plosives which were overlapped with the sonorants prior to training. Although not evident in the figure, it is mainly the voiceless plosives that show the greatest movement as we will see shortly.

It is interesting to note that while the GTM response to Spanish consonants has been affected by exposure, the same GTM's response to Chinese consonants shows very little change. Specifically, following exposure to Spanish, Chinese consonants are slightly more tightly clustered, at 98% of their separation at the outset.

5.2 Interpretation of GTM dimensions

The spatial separation or otherwise of responses to individual consonants suggests that the 3-D latent space of the GTM encodes acoustic-phonetic properties. To better explore possible interpretations of these dimensions, a series of 2-D projections of the GTM responses can be examined.

Figure 2 plots the locations of the mean centres of

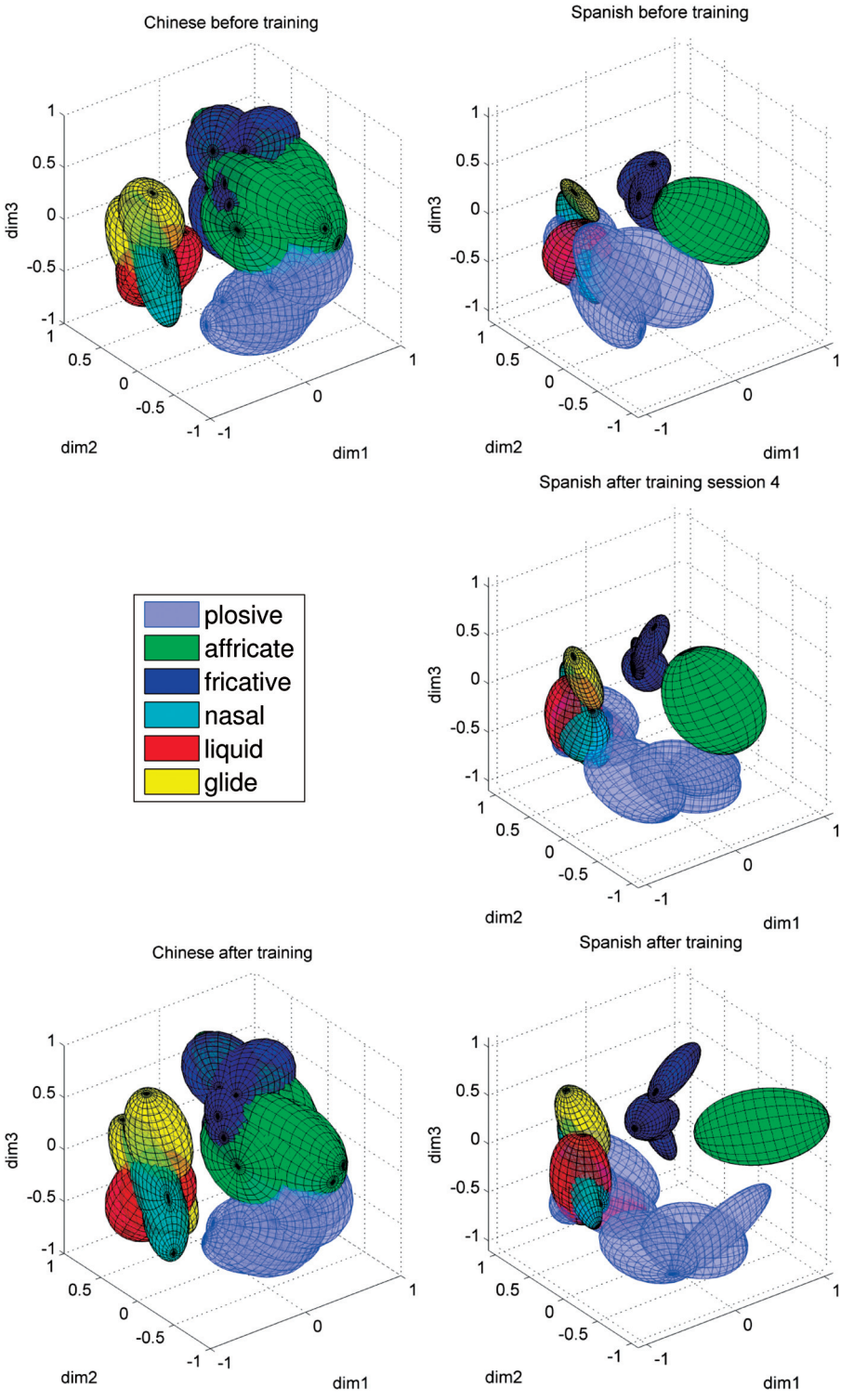

Figure 1 Spatial distribution of activation in response to L1 and L2 consonants in the GTM.

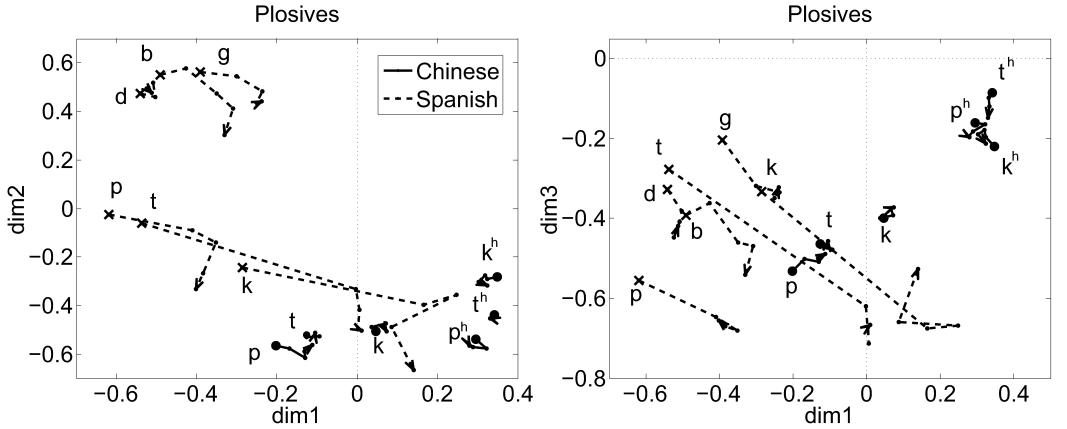

Figure 2 2-D projections of the mean locations of Chinese and Spanish plosives in the GTM before, during and after Spanish exposure. Filled circles indicate the location prior to Spanish exposure; the arrowhead shows the direction of evolution during exposure and indicates the location at the termination of exposure.

activation of Chinese and Spanish plosives projected on to dimensions 1 and 2 (left) and 1 and 3 (right) of the GTM. The arrow indicates the direction of evolution from the stage prior to Spanish exposure (indicated with filled circles), and after sessions 4, 8, 12 and 16 of exposure (indicated by changes in the gradient of the lines). Figures 3–6 show similar plots for the affricates, fricatives, nasals, and liquids/glides of the two languages.

Some of the features highlighted in the previous section are evident in these figures: the Spanish plosives are located at points in the latent space distinct from the Chinese plosives, the Spanish plosives show significant movement as a result of exposure, while the Chinese plosives are barely affected.

Considering first the projection on to dimensions 1 and 2 of the GTM (left panel of Figure 2) of the Chinese plosives, the aspirated forms /p^h, t^h, k^h/ are clearly separated from their unaspirated counterparts /p, t, k/ along dimension 1. The Spanish plosives, all of which are unaspirated, have their mean centres of response at negative values of dimension 1. This suggests that dimension 1 encodes in large part the aspiration feature. Affricates and fricatives also display various degrees of turbulent noise but this is not the case for voiced sonorants. As we can see in Figures 3 and 4, the affricates and fricatives have positive values for dimension 1, while all nasals and liquids are located at negative values (as shown in Figures 2, 5 and 6).

Figure 2 also shows that the Spanish voiced /b, d, ɡ/ and voiceless /p, t, k/ plosives are separated along dimension 2, suggesting that this dimension encodes the feature 'continuant' in the broad sense of the term which includes nasals and laterals (Mielke 2005) and reflects the fact that Spanish voiced plosives, in the intervocalic context in which they were presented in this paper, are realised as approximants. This interpretation is supported by the location of the Chinese plosives which are all voiceless and situated in the negative part of this axis.

The interpretation of dimension 3 for the plosives is less clear. The location of Chinese plosives suggests that aspiration and other noise-like speech characteristics are also reflected in this dimension, but less clearly than in dimension 1.

5.3 Effects of exposure

Exposure has some effect on the centres of activation for the Spanish plosives. While in the main they remain in the same quadrants as they occupied prior to exposure, they move towards the area occupied by Chinese plosives. Initially, the Spanish voiceless plosives /p, t, k/ were located between the Spanish voiced plosives /b, d, ɡ/ and the Chinese unaspirated plosives /p, t, k/ (Figure 2, left). It is clear from Figure 2 that, following exposure, the centres of the Spanish voiceless plosives move towards the negative half of dimension 2, becoming closer to the the Chinese unaspirated plosives. Although the Spanish voiced plosives (/b,ɡ/) also showed some movement towards the Chinese unaspirated plosives, they maintained their separation in both dimensions 1 and 2 from their Spanish voiceless counterparts, again consistent with their realisation as approximants in intervocalic contexts.

In general, the non-plosive consonants show smaller shifts as a result of exposure. However, the Spanish

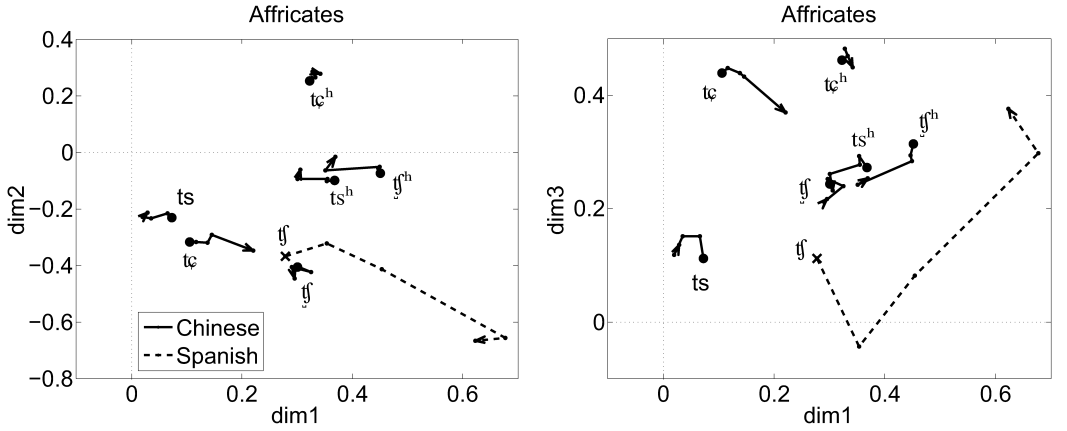

Figure 3 As Figure 2 for the affricates.

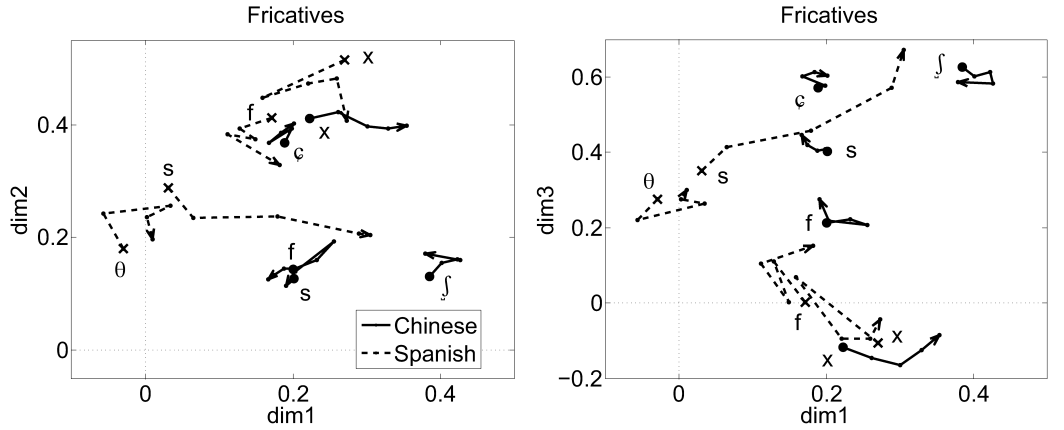

Figure 4 As Figure 2 for the fricatives.

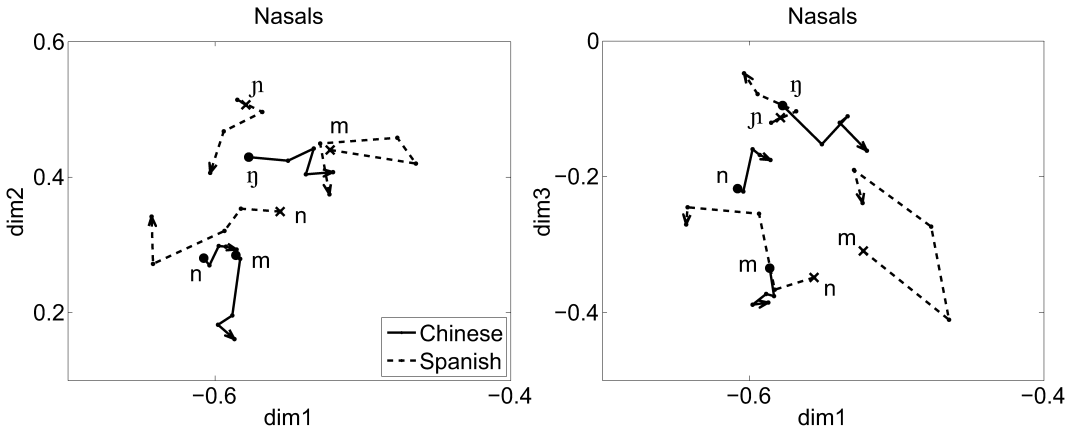

Figure 5 As Figure 2 for the nasals.

consonants /tʃ/, /s/, /l/ and /r/ did change location by a substantial amount. Spanish /tʃ/ initially belonged to the Chinese affricate cluster (Figure 3) but emerged following a similar degree of movement in the three dimensions. This is compatible with the creation of a distinct category for this sound.

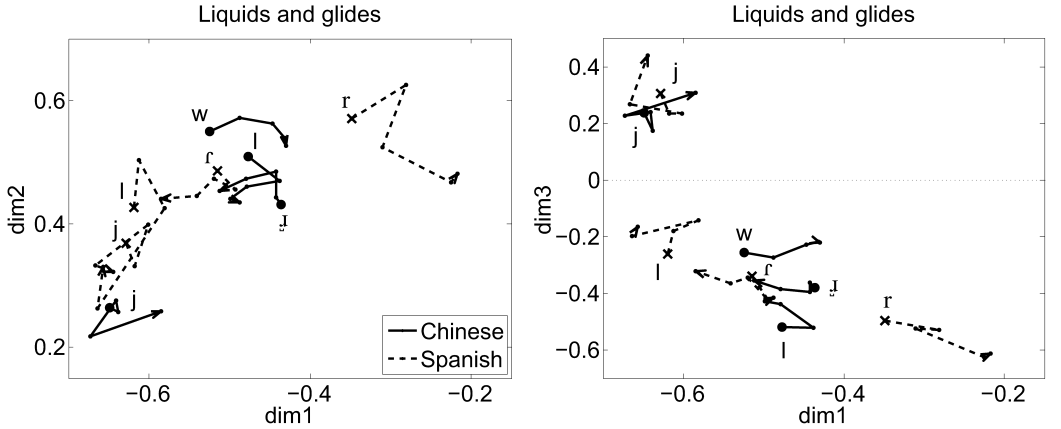

Figure 6 As Figure 2 for the liquids and glides.

In the case of Spanish /s/ (Figure 4), this sound initially overlaps partially with Chinese /s/ but is close to merging with Chinese /ʃ/. The relative locations between Chinese and Spanish consonants are to some extent consistent with the assimilation and identification results reported in the behavioural study of Gong (2013) discussed in the next section.

5.4 Relationship to behavioural data

Gong (2013) described how Chinese listeners with no previous experience of Spanish undertook an intensive training regime, identifying Spanish VCVs with feedback for incorrect responses, in 16 sessions over 4 days. Prior to and following training the same listeners carried out forced-choice identification of Spanish VCVs and performed an assimilation task where Spanish VCVs were classified in terms of Chinese categories. They also identified Chinese VCVs using an adaptive noise procedure which permitted the measurement of noise thresholds for the identification of individual consonants. Listeners were exposed to the same sequence of Spanish VCVs as used in the current study.

Gong (2013) found little evidence for attrition of Chinese categories, as measured in terms of increases in consonant reception thresholds in noise. Consistent with the results of the behavioural experiment, the distribution of Chinese consonants in the GTM did not change much after exposure to Spanish consonants.

Chinese listeners showed a marked and rapid improvement in identification of Spanish VCVs from pre- to post-test amounting to 33 percentage points on average. The increased separation of Spanish consonants in the GTM suggests that acoustic differences underpin at least some of the observed gain in classification accuracy. Much of the improvement in the behavioural study took place in the first 4 sessions, with a mean correct classification score of 46% in the pre-test rising to 74% after the fourth training session. In the GTM the location of Spanish consonants already exhibits considerable change after 4 sessions of exposure, especially for the plosives and liquids. Nevertheless, the rate of change of other consonants was more uniform as a function of exposure, suggesting that at least some of the rapid learning exhibited in the behavioural data is due to explicit feedback (encouraging, amongst other things, symbol learning) — which the GTM lacks.

The relative locations of the Chinese and Spanish consonants in the GTM reflect acoustic similarities between the two languages' sounds, and consequently might provide some explanations for the assimilation patterns in the behavioural study. For example, the Spanish voiceless plosives were located closer to the Chinese unaspirated plosives than their voiced counterparts, which is consistent with the assimilation results in Gong (2013) where the Spanish voiceless plosives showed stronger assimilations to the Chinese unaspirated plosives than did the Spanish voiced plosives. Listeners assimilated the voiceless plosives to the unaspirated Chinese ones very strongly *ab initio* while simultaneously providing category goodness ratings which indicated that they could perceive differences from their native categories, in agreement with the finding that the GTM locates the Spanish voiceless plosives at some distance from the Chinese unaspirated categories.

In the behavioural data, listener assimilations prior to Spanish exposure were very dispersed for the voiced plosives: for example, Chinese listeners categorised Spanish /g/ as one of the four different Chinese categories /k, x, l, w/, encroaching on the space of approximants and liquids (i.e. sonorants). This phenomenon

is also visible in the GTM where the plosive area encompasses the sonorant area for Spanish sounds and demonstrates that both listeners and the GTM are sensitive to the fact that these sounds are actually realised as approximants. As in the behavioural results, following exposure the locations in the GTM of the Spanish plosives are found nearer to those of the Chinese plosives.

6. Discussion

6.1 Benefits of the GTM approach

The current study demonstrates the novel application of an unsupervised learning algorithm to the problem of visualising the location of L2 consonants in an existing L1 space. The GTM approach brings a number of advantages to the modelling of L2 sound acquisition.

First, the GTM permits the use of large data corpora, enabling, for example, the model to capture some of the known within- and across-talker variability of speech tokens. Further, the quantity of speech material can be varied as we have done in order to simulate the effects of exposure.

Second, in contrast with supervised approaches such as Escudero et al. (2007) and Gong et al. (2011a), the GTM does not require a prior choice of L1 or L2 categories during learning; instead, the model can be considered as operating within the domain of acoustic similarity without the influence of category labels. Of course, we acknowledge that category labels and indeed orthography can affect category judgements (e.g. Gong et al. 2011b), but we see the GTM as a technique for exploring the pure acoustic similarity component of perceptual decision making for L2 sounds rather than as a complete model of L2 sound processing. In this sense the GTM has the potential to provide additional insights in understanding L2 speech perception.

Finally, the technique operates within a flexible statistical learning framework whose goal is to preserve similarity in the observation space and to map it to a lower-dimensional space in which visualisation is feasible. The ability to handle high-dimensional acoustic representations of speech removes the need to commit to lower-dimensional derived representations such as formant frequencies, and further enables the exploration of alternative speech feature vectors of arbitrary dimensionality such as those produced by models of the auditory periphery.

6.2 Limitations

Nevertheless, the GTM approach has a number of limitations. In practice, the mapping to latent space is limited to 3 or fewer dimensions due to the demands of visualisation. This limitation is shared with other approaches such as SOM and CDFA mentioned earlier, where a commitment to a fixed visualisation space dimensionality is required. Whether the number of dimensions is optimal for the representation and analysis of L1 and L2 sounds is not clear. A further shortcoming of the method is in the handling of temporal dynamics. In the current study we handle temporal change implicitly in two ways: the first and second time derivatives of the spectral representation are encoded, and the feature vector is sampled at three equally-spaced points in each consonant segment.

The current study did not investigate any effect of the choice of acoustic features on the form of the latent space. It is possible that features derived from more detailed simulations of perceptual processing in the auditory periphery might affect the relative locations of first and second langauge consonants. The study also employed data from male talkers only. Inclusion of female speech data might benefit from a stage of vocal tract length normalisation during feature extraction.

7. Conclusions

The current study applies generative topographic mapping to the problem of representing second language sounds in first language space in the context of Spanish and Chinese VCV sequences. By encoding acoustic-phonetic properties of L1 and L2 sounds in a common low-dimensional space, the GTM is a promising analysis tool for investigating how L1/L2 differences affect a learner's L2 perception, while the ability to vary the amount of training data enables simulation-based studies of L2 perceptual development through exposure. Using the location of Spanish intervocalic consonants in a space constructed from Chinese intervocalic consonants as a test-bed for the GTM technique, some similarities with behavioural data on assimilations is evident. Further, the GTM successfully predicts the effect of exposure to Spanish consonants by increasing the separation of clusters representing the Spanish sounds while having very little effect on existing Chinese sound clusters.

Notes

1) A posteriori probability: The probability of a hypothesis after taking into account the observed evidence.
2) *A radial basis function* is a function (e.g. a Gaussian) whose value depends solely on distance from a given point. A radial basis function network is a form of artificial neu-

ral network containing radial basis functions as activation functions.

3) *Neighbourhood function*: A mathematical function that takes inputs from neighbouring nodes in a network. For example, in the context of the Self-Organising Map (SOM) the neighbourhood function might take the form of a Gaussian that defined the variation of weights with distance from any given node.

4) *EM algorithm*: The expectation-maximization (EM) algorithm is a method for generating the maximum likehood estimate of the parameters of a statistical model. EM is an iterative algorithm that alternates between two steps. The first computes the expected value of the data likelihood function using current parameter values. The second finds new parameter estimates that maximise the expected likelihood.

5) *Forced alignment* is the process of locating specific time points in the speech signal that correspond to boundaries in a text transcription. In general alignment (e.g. during automatic speech recognition), the text is unknown. In forced alignment the text is known.

6) *Latent variables and latent space*. Latent variables are those that are not directly observed but are instead inferred via a model from observed variables. For example, latent variables might correspond to articulatory positions which have to be inferred from acoustic measurements. Latent space refers to the collective dimensions of latent variables (e.g., tongue height plus lip-rounding).

References

Best, C. T. (1995) "A direct realist view of cross-language speech perception." In W. Strange (ed.) *Speech perception and linguistic experience: Theoretical and methodological issues*, 171–204. Baltimore: York Press.

Bishop, C. M., M. Svensén and C. K. I. Williams (1998) "GTM: The generative topographic mapping." *Neural Computation* 10, 215–234.

Broomhead, D. S. and D. Lowe (1988) "Multivariable functional interpolation and adaptive networks." *Complex Systems* 2, 321–355.

de Boer, B. and P. K. Kuhl (2003) "Investigating the role of infant-directed speech with a computer model." *Acoustics Research Letters Online* 4, 129–134.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977) "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society. Series B* 39, 1–38.

Escudero, E., J. Kastelein, K. Weiand and R. van Son (2007) "Formal modelling of L1 and L2 perceptual learning: Computational linguistics versus machine learning." *Proc. Interspeech*, ISCA.

Flege, J. (2002) "Interactions between the native and second-language phonetic systems." In P. Burmeister, T. Piske and A. Rohde (eds.) *An integrated view of language development: Papers in honor of henning wode*, 217–244. Trier: Wissenschaftlicher Verlag.

Flege, J. E. (1995) "Second-language speech learning: Theory, findings, and problems." In W. Strange (ed.) *Speech perception and linguistic experience: Theoretical and methodological issues in cross-language speech research*, 233–277. Baltimore: York Press.

Gong, J. (2013) *Computational modelling of sound perception in a second language.* PhD thesis, University of the Basque Country.

Gong, J., M. Cooke and M. L. García Lecumberri (2011a) "A computational modelling approach to the development of L2 sound acquisition." *Proc. International Congress of Phonetic Sciences*, 755–758.

Gong, J., M. Cooke and M. L. García Lecumberri (2011b) "Towards a quantitative model of Mandarin Chinese perception of English consonants." In M. Wrembel, M. Kul and K. Dziubalska-Kolaczyk (eds.) *Achievements and perspectives of SLA of speech*, 103–113. Oxford: Peter Lang.

Guenther, F. H. and M. N. Gjaja (1996) "The Perceptual Magnet Effect as an emergent property of neural map formation." *Journal of the Acoustical Society of America* 100, 1111–1121.

Hebb, D. O. (1949) *The organization of behavior: A neuropsychological theory*. New York: Wiley and Sons.

Iverson, P. and P. K. Kuhl (1996) "Influences of phonetic identification and category goodness on American listeners' perception of /r/and/l/." *Journal of the Acoustical Society of America* 99, 1130–1140.

Johnson, D. E. (1998) *Applied multivariate methods for data analysis*. Pacific Grove, CA: Duxbury.

Klecka, W. R. (1980) *Discriminant analysis*. Sage Publications.

Kohonen, T. (1988) "The "neural" phonetic typewriter." *Computer* 21, 11–22.

Kuhl, P. K. (1993) "Innate predispositions and the effects of experience in speech perception: The native language magnet theory." In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage and J. Morton (eds.) *Developmental neurocognition: Speech and face processing in the first year of life*, 259–274. Berlin: Springer.

Mielke, J. (2005) "Ambivalence and ambiguity in laterals and nasals." *Phonology* 22, 169–203.

Miyazawa, K., H. Kikuchi and R. Mazuka (2010) "Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model." *11th Interspeech*.

Morrison, G. S. (2006) *L1 & L2 production and perception of English and Spanish vowels: A statistical modelling approach*. PhD thesis, University of Alberta.

Nabney, I. T. (2004) *NETLAB: Algorithms for pattern recognition*. Springer.

Salminen, N. H., H. Tiitinen and P. J. May (2009) "Modeling the categorical perception of speech sounds: A step toward biological plausibility." *Cognitive, Affective, & Behavioral Neuroscience* 9, 304–313.

Strange, W., O.-S. Bohn, S. A. Trent and K. Nishi (2004) "Acoustic and perceptual similarity of North German and American English vowels." *Journal of the Acoustical Society of America* 115, 1791–1807.

Tatsuoka, M. M. (1970) *Discriminant analysis: The study of group differences*. Champaign IL: Institute for Personality and Ability Testing.

Thomson, R. I., T. M. Nearey and T. M. Derwing (2009) "A modified statistical pattern recognition approach to measuring the cross-linguistic similarity of Mandarin and English vowels." *Journal of the Acoustical Society of America* 126(3), 1447–1460.

Vallabha, G. K., J. L. McClelland, F. Pons, J. F. Werker and S. Amano (2007) "Unsupervised learning of vowel categories from infant-directed speech." *Proceedings of the National Academy of Sciences* 104, 13273–13278.

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

(Recieved 0.0, 0, Accepted 0.0,0)