Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 69 (2015) 17–30

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A quantitative model of first language influence in second language consonant learning

Jian Gong [a,b,*], Martin Cooke [c,b], Maria Luisa García Lecumberri [b]

[a] School of Foreign Languages, Jiangsu University of Science and Technology, 2 Mengxi Road, 212003 Zhenjiang, China
[b] Language and Speech Laboratory, Universidad del País Vasco, Paseo de la Universidad 5, 01006 Vitoria, Spain
[c] Ikerbasque (Basque Science Foundation), Spain

Received 2 July 2013; received in revised form 13 January 2015; accepted 5 February 2015
Available online 14 February 2015

Abstract

Theoretical models argue that listeners' perception of second language sounds is heavily influenced by their native language phonology, a prediction borne out by behavioural studies. However, we lack quantitative models capable of making more precise predictions of the way in which the first and second language sound systems interact. The current study introduces a computational modelling framework that permits comparison of different second language learning strategies which vary both in the degree of first language influence as well as in the manner in which second language input is combined with existing first language knowledge. Six different model variants were evaluated by comparison with behavioural data on a task involving the identification of intervocalic consonants of Castilian Spanish by Mandarin Chinese listeners. All approaches demonstrated a similar pattern of rapid improvement with exposure to that observed in listeners. However, approaches that made use of independent first and second language models made the best predictions. An approach that excluded first language influence both predicted lower listener identification levels in the initial stages of learning and higher scores in later stages, demonstrating that first language experience helps to bootstrap second language sound learning but ultimately hinders identification. However, modelling outcomes also demonstrate that no single approach can account for the identification patterns for all consonants, suggesting that learners deploy different approaches to the learning of individual sounds.

Keywords: Foreign language sound acquisition; Consonant identification; Rapid learning; Computer model

1. Introduction

Acquiring the sounds of a new language as an adult is different in at least one crucial respect from the linguistic situation that confronts us as infants: as adults, we already possess a well-developed phonological system. Investigation of how the first (L1) and second (L2) language sound systems interact to influence the development of phonetic competence in a second language is an active area of study, and the degree to which adult learners benefit – or do not – from prior experience remains an important issue both in broadening our understanding of second language learning and in the study of general phonological representations and processes.

Besides the age of acquisition, the influence of the first language is probably the single strongest factor in non-native sound acquisition. Theoretical models (e.g., Kuhl, 1993; Best, 1995; Flege, 1995) agree that perceived similarities between native and non-native phonetic categories play a crucial role in non-native sound perception. Learners may process non-native sounds in terms of their L1

* Corresponding author at: School of Foreign Languages, Jiangsu University of Science and Technology, 2 Mengxi Road, 212003 Zhenjiang, China.
E-mail addresses: habaogj@hotmail.com (J. Gong), m.cooke@ikerbasque.org (M. Cooke), garcia.lecumberri@ehu.es (M.L. García Lecumberri).

http://dx.doi.org/10.1016/j.specom.2015.02.003
0167-6393/

categories given sufficient similarity between the two in a process variously known as 'equivalence classification' (Flege, 1995) or 'perceptual assimilation' (Best, 1995), which can prevent category formation for non-native sounds. Best and Tyler (2007) suggest that L2 learning is a process of fine-tuning and possibly 're-phonologizing', in which learners might stretch or modify their existing L1 categories to accommodate L2 perception (see also Bundgaard-Nielsen et al., 2011).

While theoretical models have provided valuable hypotheses and insights concerning the influence of the L1 phonological system on the development of non-native sound perception and production, what these models lack is the ability to predict in a quantitative manner how the established L1 system and the evolving L2 system interact. For instance, it is difficult to answer questions about the nature of the relationship between the amount of exposure to second language sounds and the rate at which L2 category identification improves.

Computational simulations provide a useful adjunct to theoretical models, and have been used in studies of L1 sound acquisition to show how infants learn native vowel categories and to simulate the perceptual magnet effect (de Boer and Kuhl, 2003; Vallabha et al., 2007; Lake et al., 2009; McMurray et al., 2009). Statistical pattern recognition approaches have also been applied recently to simulate human perceptual assimilation tasks in order to measure cross-language category similarities directly from the acoustic data, providing quantitative formal evaluation of the predictions of theoretical models (Strange et al., 2004; Morrison, 2009; Thomson et al., 2009; Gong et al., 2010). Computational approaches have also been used in studies of the development of L2 perception. For example, in Escudero et al. (2007), machine learning and computational linguistic models were used to simulate and visualise the evolution of learners' L2 vowel spaces. Hidden Markov modelling (HMM) techniques were adopted by Gong et al. (2011) in a modelling study investigating the effect of different ratios of L1 and L2 exposure in identifying second language. These studies not only demonstrated the promise of using computational approaches to model the L2 learning process, but also highlighted a key advantage of simulation, viz. the ability to contrast and evaluate competing models while maintaining control over factors such as the degree and type of exposure. However, simulation studies to date have been somewhat limited in scope. Many existing models have been constructed using either synthetic speech (Escudero et al., 2007) or simplified and abstract speech parameters (e.g., F1/F2 values or VOTs) (de Boer and Kuhl, 2003; Vallabha et al., 2007; Strange et al., 2004; Thomson et al., 2009) while small subsets of vowels or consonants have usually been chosen as the modelling targets (de Boer and Kuhl, 2003; Vallabha et al., 2007; Morrison, 2009; Escudero et al., 2007).

The current study addresses the issue of how L1 knowledge interacts with L2 exposure at the outset of second language learning. We do so by evaluating how closely a number of computational models predict findings from a recent study of non-native consonant identification (Gong, 2013). In that study (reviewed in Section 2 below) Chinese listeners with no experience in Spanish took part in an intensive high-variability perceptual training programme of the kind found to be effective in earlier studies (e.g., Logan et al., 1991; Lively et al., 1993; Bradlow et al., 1997). Listeners were required to identify Spanish consonants drawn from the full consonant inventory, when presented in intervocalic context. Using Gong (2013) as the behavioural reference, in the current study we apply computational modelling to investigate the development of second language learning when confronted by an extensive L2 sound inventory. Our models differ in the manner in which speech material in the L1 and L2 interact during learning and consonant recognition. These models use precisely the same training data, in the same sequence, as made available to listeners in the behavioural study. As such, listeners and models had identical exposure to the consonants of the target language during training. As in Gong et al. (2011), HMMs were used to represent sound categories. An initial HMM set was trained using Chinese data and consonant categories, and subsequently retrained for Spanish consonant categories based on listeners' assimilations of Spanish sounds.

One model – BLEND – is motivated by the hypothesis that it is the raw amount of L2 exposure that is the key determinant of listeners' identification of L2 sounds. BLEND operates by adding in progressively larger quantities of L2 stimuli and re-learning HMM parameters *ab initio*. A second model, ADAPT, is based on the idea that it is not just quantity but the sequence of exposure to new sounds that matters. Rather than re-training at each stage of learning, HMM parameters are adapted using Bayesian speaker adaptation techniques. A further model, SEPARATE, represents the hypothesis that learners of a new sound system are capable of maintaining separate L1 and L2 representations and that they use only the latter to identify L2 speech sounds. We additionally evaluate versions of each of the BLEND, ADAPT and SEPARATE approaches in which HMMs for the L1 are activated in parallel. These models – PAR-BLEND PAR-ADAPT and PAR-SEPARATE – represent L1/L2 interaction at the level of categories.

Section 2 reviews the behavioural study of Gong (2013) and describes the Spanish and Chinese speech materials used in the current study. Section 3 describes how listener assimilation results inform the development of the initial model set, while the six modelling approaches are explained in Section 4. Simulation results are presented and compared to listeners in Section 5.

2. Behavioural study

The stimuli and human baselines used in the current study are described in Gong (2013). Here, we review the tasks, speech materials, and outcomes of that study.

2.1. Listeners

A cohort of 20 native Mandarin Chinese listeners (11 females and 9 males) aged between 20 and 23 (mean: 20.5 years) took part in an intensive seven-day perceptual training regime. All listeners were from north China and had a northern Mandarin dialect. In this dialect, the feature "retroflex" is important in distinguishing some affricate and fricative contrasts (e.g., denti-alveolar /s/ and retroflex post-alveolar /ʃ/). All listeners were studying medical courses and none had studied Spanish or any other Romance language before, and no listener had lived outside China. An additional group of eight listeners served as a control. They had a similar profile to those in the main study and undertook only the pre- and mid-tests with a gap of two days, but received no training in the interim.

2.2. Tasks

On days 1, 4 and 7, listeners identified Spanish consonants in intervocalic context (VCVs) using an 18-alternative forced choice paradigm. The task and materials on each of the three test days (which we will refer to as pre-test, mid-test and post-test) were identical. A total of 360 VCVs were used during testing, composed of 10 examples of both initially and finally-stressed tokens for each of 18 Spanish consonants.

On the remaining four days (days 2, 3, 5, 6) listeners underwent perceptual training on VCV tokens that differed from those used during the testing phases. Training sessions had a similar format to the test sessions except that participants received immediate feedback on incorrect responses. For such responses, a button with the correct answer was highlighted and activated, and listeners were then required to listen to the stimulus exactly once before continuing to the next token. Participants took part in 4 training sessions on each day. Each training session contained 180 tokens, 10 for each of 18 Spanish consonants (-five each for initial and final stress). Nine types of vowel context were used in order to increase exposure to co-articulatory variation. No token was repeated during the 16 training sessions, leading to a total exposure to 2880 different tokens, 160 for each consonant.

In addition to the L2 consonant identification task, participants also carried out a category assimilation test on each of the test days. This task involved mapping Spanish VCVs to Chinese consonant categories. As in the identification test, a total of 360 Spanish VCVs were categorised along with 48 Chinese VCV tokens which served as control items. The results from the assimilation pre-test were used in the production of the initial set of Spanish VCV models, as described in Section 3 below.

2.3. Speech materials

Naturally-produced VCV sequences for all 324 combinations of the 18 Spanish consonants /p, b, t, d, k, g, ʧ, f, θ, s, x, m, n, ɲ, l, ɾ, r, j/ and vowels /a, i, u/, with initial and final vowel stress, served as stimuli for the training and test phases of the study. Sixteen native male Spanish talkers produced the full VCV set. These talkers originated from the Basque Country, in northern Spain. Like most northern and central peninsular varieties, their accent has a phonological contrast between the interdental fricative /θ/ vs. sibilant /s/. As is increasingly the case in Spanish, these speakers do not produce the palatal lateral /ʎ/ corresponding to orthographic "ll". Instead, there is a neutralization between this phoneme and the palatal continuant. Whether this continuant phoneme is an approximant, a fricative, an affricate or even a plosive, is a much debated topic (Quilis, 1997; Martínez-Celdrán et al., 2003; Hualde, 2005; Fernández, 2007). All these realisations are possible, depending on phonetic context, style and regional variation. Material from 10 talkers was used for model training while VCVs from the remaining 6 talkers were used in model testing. Chinese control tokens for the assimilation tests came from a VCV corpus collected for an earlier English-Chinese mapping study (Gong et al., 2010). The Chinese VCV corpus was also used in the current study to train the initial set of Chinese models (see Section 3). The Chinese corpus contains tokens for all 24 Chinese consonants /p^h, p, t^h, t, k^h, k, ts^h, ts, $ʧ^h$, ʧ, $ʨ^h$, ʨ, f, s, ʃ, ɕ, x, m, n, ŋ, ɻ, ɹ, l, j, w/ in the same 9 vowel contexts as used for the Spanish corpus. Since lexical stress is not a feature of Chinese, the corpus did not differentiate the two stress types. The corpus contains speech material provided by 17 native male speakers of Mandarin.

2.4. Results

Listener scores during training and testing are visualised alongside those from the models in subsequent sections. Fig. 2 depicts per-consonant identification rates prior to training, while consonant confusions at the same point are provided in the upper panel of Fig. 3. Identification scores in the pre-test and after each session of training are shown in Fig. 4.

Prior to training, mean identification performance was 46%, a level well above chance, demonstrating the role of L1 experience in L2 consonant identification even for naïve learners. Learners exhibited extremely rapid improvement in the early stages of training, reaching a rate of nearly 75% correct identification after 4 short training sessions, illustrating a considerable capacity for the formation of new L2 categories. By comparison, the identification rate for the control group did not change significantly between pre- and mid-test (44% vs 46%; $p = 0.18$), confirming that the improvement seen in the experimental group was due to the feedback they received during the training phase.

The improvement was notably smaller for voiced plosives. L2 acquisition of plosive VOT and, in the case of voiced plosives, of their approximant (spirantized) realizations such as the ones in our corpus, has been reported to

be particularly problematic, at least for production by L1 English speakers (Díaz-Campos, 2004; Díaz-Campos, 2014; Zampini, 1994; Zampini, 2014) and L1 Chinese speakers (Chen, 2007). Our results extend to perception the finding that spirantized plosives are problematic for these learners.

Subsequent sections describe model construction and evaluation. In all cases, the speech material used for training and testing the models was precisely that used in the behavioural study.

3. Initial models

3.1. HMM architecture

All models to be introduced in Section 4 are based on continuous density HMMs, trained using the HTK toolkit (Young et al., 2006) using 39-component vectors composed of 12 mel-frequency cepstral coefficients plus energy, and their first and second temporal derivatives, computed every 10 ms. Individual vowels and consonants are modelled as 3-state HMMs and combined during recognition into VCVs. Within each state, a mixture of Gaussian distributions represents speech observations deemed to belong to that state. A limit of 10 mixture components per state was determined in pilot tests as the best tradeoff between model accuracy and use of the available training data.

3.2. Mapping Mandarin Chinese to Spanish

An initial set of Chinese HMMs was constructed using the procedure described above. However, in order to compare models with listeners undertaking the L2 sound categorisation task, it is necessary for the Chinese models to provide responses in terms of *Spanish* sound labels. Since the Chinese-Spanish sound mapping is not one-to-one – indeed, the sizes of their consonant inventories differ – satisfying this requirement is potentially problematic. For Chinese listeners as a cohort the Chinese-Spanish mapping is many-to-many: Spanish exemplars with the same consonant label can map on to different L1 categories, and tokens with different labels can map on to the same category.

One possible solution is to map Chinese consonant models to the closest Spanish label as determined, for example, by the HMM which most frequently 'assimilates' to the Spanish sound (i.e., the model which produces the highest recognition score for a given Spanish consonant, as measured over a test corpus). Listener assimilations from Gong (2013) are reproduced in Fig. 1. These demonstrate, for example, that while Spanish /p/ is predominantly categorised by native Chinese listeners as Chinese /p/, Spanish /b/ also assimilates to Chinese /p/ and to a greater extent Chinese /w/. Several drawbacks of the label mapping approach are evident from this figure. First, there may be more than one Chinese target for a Spanish category with similar assimilation frequencies – as is the case for

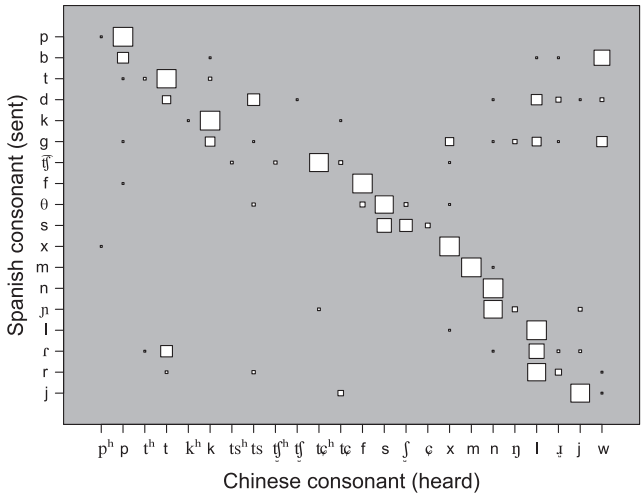

Fig. 1. Visualisation of listener assimilation responses. The area of each square is proportional to the number of assimilations of the Spanish consonant to the Chinese category.

Spanish /b/ – yet once the single Chinese model has been chosen, these other categories play no further role in the modelling process. A second issue is that duplicate Chinese models can be chosen for the initial Spanish-labelled model set, due to the same Chinese sound being the most frequent assimilation target for multiple Spanish categories. This is the case for Chinese /l/, which is the most frequent assimilation for Spanish /ɾ/ and /ɾ/ as well as Spanish /l/. Having the same initial model is problematic in subsequent consonant identification simulations since several models will share precisely the same fit.

Here, these problems are resolved using a different approach, which we call 'assimilation-based model retraining'. First, according to the listener-derived assimilation percentages from each individual Spanish sound to different target Chinese categories, VCV tokens in the same proportions are chosen at random from these Chinese categories to form a training set for that Spanish sound. Subsequently, a new HMM labelled with the corresponding Spanish sound is built from this training set. For example, Spanish /s/ was assimilated to the Chinese sounds /s, ʃ, ɕ/ with assimilation percentages 53%, 41% and 6% respectively, so a new HMM labelled as Spanish /s/ is trained on a mixture of data with 53% from the Chinese /s/ category, 41% from Chinese /ʃ/, and 6% from Chinese /ɕ/. To ensure balanced exposure across the different Spanish categories, the total number of VCV tokens for each new HMM is fixed at a constant value (set here to 120, which is the average number of tokens across the Chinese categories). As a consequence of assimilation-based model retraining, all individual HMMs are unique and represent not just the single most frequent assimilation target but the distribution of targets. Retrained models form the starting point for most of the modelling strategies, and are referred to in subsequent sections as the 'Initial Model' (IM) set.

3.3. Initial model set performance

As noted above, listeners identified around 46% of Spanish consonants correctly prior to training. The initial model set derived by assimilation-based model retraining results in a somewhat higher identification rate of 53%. Fig. 2 compares per-consonant identification scores of listeners and the initial model set. Consonant confusion matrices for listeners and the initial model set are provided in Fig. 3.

Apart from the voiceless plosives, most of the consonants have relatively similar identification rates in the pre-test for listeners and the initial models. Listeners – but not the initial models – almost universally mis-categorised the voiceless plosives as their voiced equivalents. Other listener-model disparities can be seen for /θ/ and /n/. As shown by the confusion matrices prior to training (Fig. 3), listeners largely mis-categorised /θ/ as /s/ whereas the model produced the more acoustically-based confusion of /θ/ and /f/. While listeners found /n/ unproblematic, the initial model had difficulty in distinguishing it from /ɲ/. In fact, due to the listener assimilations used in the initial model construction procedure, the initial /n/ and /ɲ/ models were both trained on a large proportion of Chinese /n/ data and hence were similar at the outset, giving rise to many mutual confusions in the pre-test.

3.4. Interim discussion

Clearly, listeners are able to respond to L2 sounds in terms of L1 categories, albeit with a greater or lesser degree of confidence or goodness of fit. We argue that in order to model category identification it is important to reflect these biases in our models. However, our goal here is purely pragmatic: we require a mechanism capable of providing category judgements in the absence of L2 data. The way in which behavioural assimilation data is used in constructing the initial model set should not be understood as a hypothesis about how listeners create new categories, a process which is outside the scope of the current modelling study.

Listeners' mis-categorisation of Spanish voiceless plosives as their voiced equivalents in the pre-test is likely to be due to orthographic influences from Pinyin, a system that most Chinese children learn in primary school to help them to remember the pronunciation of Chinese characters. The Pinyin system uses Roman letters to mark the sounds. The letters 'b', 'd' and 'g' are the graphemes used

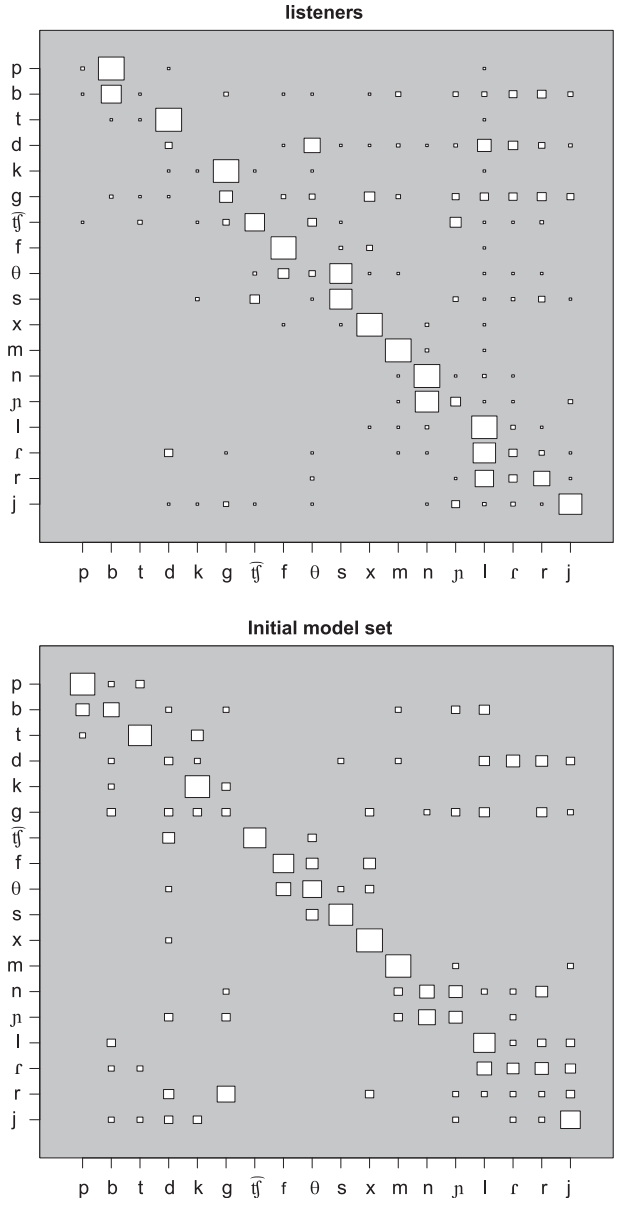

Fig. 3. Visualisation of consonant confusions prior to training. Row labels indicate stimuli while column labels indicate responses. Top: listener confusions. Bottom: initial model confusions.

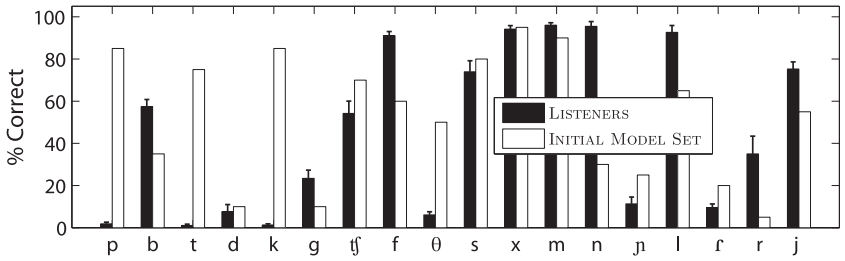

Fig. 2. Model-listener comparison of consonant identification rates prior to training. Error bars indicate 1 standard error.

in the Pinyin system to represent the Chinese voiceless unaspirated counterparts of Spanish /p, t, k/. Orthographic transfer is likely to account for their low rates of identification. Here, the modelling technique of assimilation-based retraining ensures that the acoustic properties of the corresponding Spanish sounds are well represented. However, the models are unable to take account of orthographic transfer, leading to listener-model disparities for these sounds.

4. Models of L2 consonant identification

Six alternative models were developed to explore the role of L1 influence in early-stage L2 consonant identification. In this section we motivate these models and detail each approach. The differing use of speech data and models during training and testing amongst the strategies is summarised in Table 1.

4.1. Blend

The Blend model mixes different proportions of Chinese and Spanish training data to simulate the exposure to L2 sounds available at different stages in the learning process. For each Spanish category *C* at training session *i*, Spanish data from all training sessions up to and including *i* for that category is added into the (fixed) Chinese training data used in assimilation-based model training to construct the initial model for category *C*. This blended set of Spanish and Chinese VCVs is then used to train the model for *C* after session *i*.

For instance, as explained in Section 3.2, in the initial model set the Spanish /s/ model is derived using 120 Chinese VCV tokens drawn in various proportions from the categories /s, ʃ, ɕ/. At the end of the first training session, the blended /s/ model is trained on the mixture of these 120 Chinese tokens *plus* the 10 Spanish /s/ tokens heard by listeners during the first session. Similarly, in the second session the original 120 Chinese tokens are joined by 20 Spanish /s/ exemplars, and in subsequent training stages additional blocks of Spanish /s/ tokens are added to the training data set used in the previous stages. The Blend modelling approach represents a possible learning mode based on gradual merging of new L2 tokens with prior L1 knowledge. As more L2 tokens are added, the influence of the L1 is expected to be reduced.

4.2. Adapt

The Adapt approach is similar to Blend but uses model adaptation techniques developed for automatic speech recognition to progressively modify the initial HMMs using Spanish speech material available up to the given training session. The key difference between the Adapt and Blend approaches is that Adapt maintains the sequential order in which additional data is used, while Blend retrains the models from scratch at each stage, removing any sequential information. The Adapt approach better reflects a listener's exposure and sequential learning, and allows us to measure the importance of sequential ordering effects (e.g., recency of experience) in identifying non-native language sounds.

Adapt uses the maximum *a posteriori*(MAP) approach. MAP is a Bayesian speaker adaptation technique which treats the existing HMM parameters as prior information and uses a re-estimation procedure to find parameter values that maximize the model's *a posteriori* probability of generating the adaptation target data (Gauvain and Lee, 1991; Gauvain and Lee, 1994). MAP adaptation is implemented in a cascaded fashion as follows. In each training stage, the model trained at the previous stage is used as the current model. The parameters of this model are then modified based on the new block of Spanish speech material used at the equivalent stage of listener training. MAP is a *supervised* adaptation technique: since the Spanish training material has category labels, adaptation for each category is applied to the corresponding HMM.

Once again taking the /s/ model as an example, at the end of first training session the parameters of the initial /s/ model are modified to adapt to the 10 Spanish /s/ tokens heard in the first session. In each subsequent training session the parameters of the adapted /s/ model will be further modified based on the data heard in that session.

4.3. The Separate model

A further model – Separate– was constructed using only the Spanish speech material available at each point in the training procedure. For example, the Separate /s/

Table 1
Model and data requirements for each of the modelling strategies. CH: Chinese data; IM: initial HMM set; SP_i: Spanish data from training session *i*; BM_i, AM_i, SPM_i: Blend, Adapt and Separate HMM set after training session *i*.

Strategy	Session	Source model	Training data	Adaptation data	Output/test model	Test model (parallel)
Blend	1	–	CH+SP_1	–	BM_1	BM_1 + IM
	n	–	CH+$SP_1 \ldots SP_n$	–	BM_n	BM_n + IM
Adapt	1	IM	–	SP_1	AM_1	AM_1 + IM
	n	AM_{n-1}	–	SP_n	AM_n	AM_n + IM
Separate	1	–	SP_1	–	SPM_1	SPM_1 + IM
	n	–	SP_n	–	SPM_n	SPM_n + IM

model is initially learnt from the 10 Spanish tokens presented in the first session. The second SEPARATE /s/ model is trained on the 10 tokens from the first session plus the 10 tokens from the second session. In this way, at the end of the 16th training session, the SEPARATE /s/ model is trained on all 160 Spanish /s/ tokens used in the behavioural experiment. This model serves to measure the extent to which the Spanish data alone are sufficient to permit consonant identification after each training session, and as such represents the potential performance for a hypothetical new language learner who is not subject to any influence from Chinese L1 models, or, equivalently, is capable of keeping non-L1 sounds apart from their internal L1 sound system. Additionally, the SEPARATE model enables us to estimate the effects of data paucity during early stages of exposure.

4.4. Parallel L1/L2 activation

The BLEND and ADAPT approaches model the interaction of L1 and L2 speech data at an acoustic feature level, from which a *single* HMM is built for each category. Thus, the initial Chinese models are modified thoughout the simulated training programme and consequently the original L1 influence is progressively weakened. To explore the effect of maintaining *separate* L1 models, parallel versions of BLEND, ADAPT and SEPARATE – denoted PAR-BLEND, PAR-ADAPT and PAR-SEPARATE – were constructed. Parallel models represent L1–L2 interaction at the categorical rather than the acoustic level.

In practice, application of the parallel approach differs from its counterpart non-parallel approach only in the test phase. The initial model set is evaluated alongside the model sets created by the BLEND, ADAPT or SEPARATE approaches. Hence, for each Spanish consonant there are two HMMs with the same consonant label. For example, in the case of the PAR-SEPARATE strategy there is an HMM labelled with category *C* from the SEPARATE model set and another labelled with category *C* from the initial model set. In the decision process, log likelihoods from these pairs of models are summed, and the most likely category chosen as the parallel model's response.

5. Simulation results

After each training session, identification tests were carried out using models trained under the different modelling strategies, using precisely the same test data presented to listeners. This section compares the six modelling approaches outlined above with listener scores, both overall and for individual Spanish consonants.

5.1. Overall identification performance

The upper panel of Fig. 4 plots identification scores as means over all 18 Spanish consonants through the 16 training sessions for listeners and for each of the six modelling strategies. Scores in the pre-test using the initial model set are also plotted.

The improvement in identification scores as a result of training observed in listeners is broadly reflected in the different modelling strategies, with a rapid increase over the first 2–4 sessions followed by a more gradual increase. All modelling strategies asymptoted at an identification rate within 5 percentage points of that obtained by listeners. The main differences between the modelling approaches can be seen in the detailed evolution of scores during training. The lower panel of Fig. 4 shows the difference between model and listener scores for each strategy. Both ADAPT and PAR-ADAPT produce higher scores than listeners throughout the entire training process, although they converge to the behavioural data in the latter stages of training. The evolution of both BLEND and PAR-BLEND relative to human listeners is more varied in the initial stages. In the latter stages, from training session 10 onwards, the patterns for BLEND and ADAPT and their parallel versions are quite similar. Indeed, while differences exist, in the main the parallel versions are quite similar to the non-parallel counterparts for both of these modelling strategies.

This is not the case for the SEPARATE and PAR-SEPARATE approaches, which show rather different patterns both from each other and from the other strategies during the training simulation. Unlike ADAPT and BLEND, the SEPARATE model predicts substantially lower listener identification scores in the very early stages, but predicts higher scores from training session 5 onwards. The PAR-SEPARATE approach shows more variability, but generally predicts lower scores for the first half of the training period and marginally higher scores thereafter. It is interesting to note that the identification rate for models trained on the Spanish speech material available in the first training session alone already reaches about 47%. While quite high, this figure is substantially lower than the 59% obtained by listeners, suggesting that the latter group benefit from their L1 experience in recognising Spanish tokens in the early stages of exposure. However, during later stages this same L1 experience appears to count against listeners since the pure Spanish models out-perform listeners after a certain amount of initial exposure.

To better appreciate the differences between the models, Table 2 quantifies both the root mean square (RMS) and the standard deviation of the difference scores across training sessions. The latter reflects how well the shape of the performance versus training epoch curve matches that of listeners. The PAR-BLEND model provides the best fit to the overall distance while the PAR-ADAPT model provides the best fit to the overall shape. The pure Spanish model has the poorest fit for both metrics. Parallel versions always out-perform their non-parallel counterparts. Incorporation of parallel models provides the largest benefit for the SEPARATE model, but has relatively little effect on BLEND and ADAPT. This result probably reflects differences in the

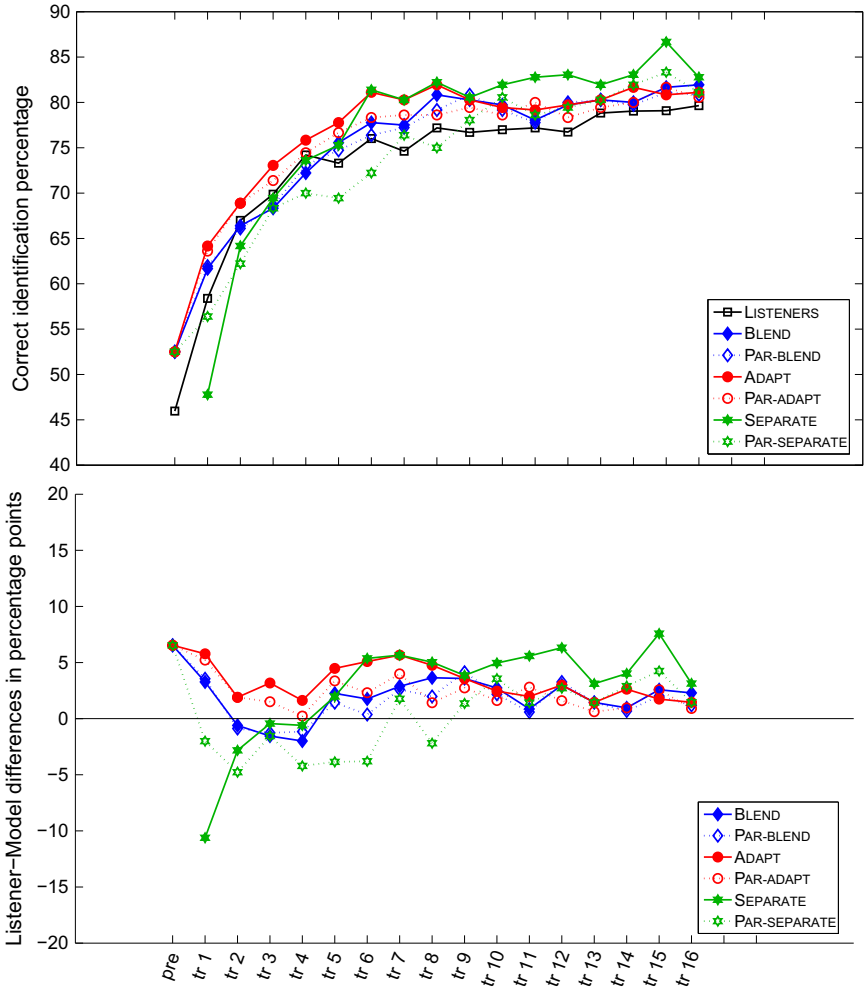

Fig. 4. Consonant identification scores as a function of training session for listeners and models. Scores for the initial models in the pre-test are also shown ('pre'). Top: absolute scores. Bottom: model-listener differences.

Table 2
Model-listener RMS distance and shape indices (percentage points).

Model	BLEND	PAR-BLEND	ADAPT	PAR-ADAPT	SEPARATE	PAR-SEPARATE
RMS	2.4	2.1	3.5	2.5	5.1	2.9
shape	1.8	1.6	1.5	1.3	4.5	3.0

Chinese 'exposure' that the different models possess, as discussed below.

For presentational clarity in subsequent analyses of individual consonants, we omit BLEND and ADAPT since they performed less well than their parallel counterparts. However, we continue to plot the SEPARATE approach since it is significantly different from its parallel counterpart in that SEPARATE is the only one of the six strategies which exploits solely L2 data.

5.2. Individual consonant identification during testing

Fig. 5 presents a more detailed comparison of model and listener scores at the mid- and post-test stages for each individual consonant. Consonant confusion matrices corresponding to the post-test for listeners and four models are shown in Fig. 6.

By the time of the mid-test (upper panel of Fig. 5) many models show an improved correspondence with listeners. Apart from the sounds /p, θ, ɾ/ one or more models are within a few percentage points of listener identification rates. However, no single model produces good predictions of listener scores for all consonants. Indeed, for the four models plotted in Fig. 5, model-listener correlations are in the range [0.63–0.77]. This result raises the possibility that different consonants are subject to different processes best described by one or other of the models.

The outcome is broadly similar in the post-test (lower panel of Fig. 5), showing some clear model-listener similarities. For example, the voiced plosives /b, d, g/ were

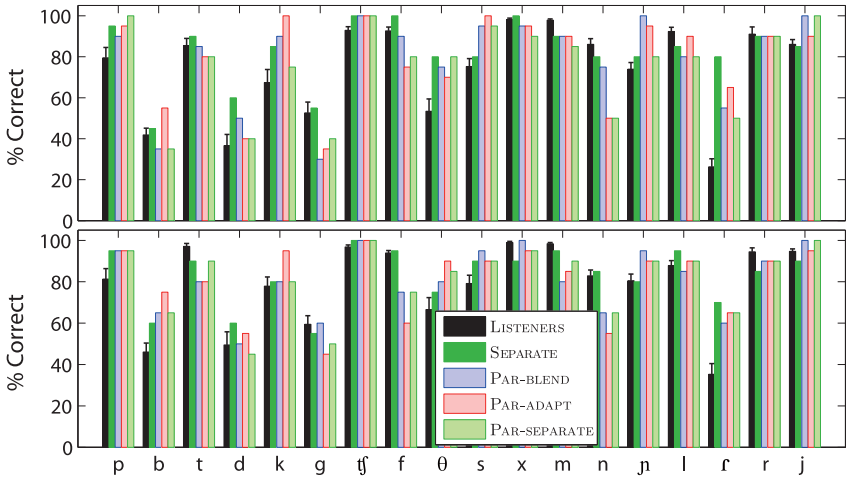

Fig. 5. Model-listener comparison of consonant identification rates at the mid-test (top) and post-test (bottom) stage. Error bars indicate 1 standard error.

amongst the least accurately identified by listeners, probably due to the lenited character of these sounds in intervocalic positions which are quite different from Chinese plosive categories. Listeners' identifications of these sounds, particularly in the pretest, show that they are aware of their continuant nature. Indeed, although listeners do confuse voiced plosives with their voiceless counterparts, there is a degree of dispersion in the identifications of these voiced plosives which demonstrates listeners' uncertainty and results in non-plosive choices. For instance, after training /d/ is still heard 28% of the time as a non-plosive such as one of /θ, l, ɾ/. Like the learners, our models show much worse performance for voiced "plosives" than for voiceless plosives, with a considerable amount of dispersion in the confusions. A noticeable difference is that listeners' main confusions are still the voiceless counterparts of the plosives (Fig. 6), whereas for the models other voiced plosives are chosen just as frequently, indicating a greater weighting of the voice feature by the models. For voiced plosives, all models over-predict identification rates for /b/, while for the other two plosives there is a mixed picture: PAR-BLEND is closest for /d, g/ but the remaining models suggest lower identification rates for /g/. Similarly, in the case of voiceless plosives, all models over-predict for /b/ and under-predict for /t/. There are also some differences between models and listeners for voiceless plosives. Even after training listeners continue to confuse them with their voiced counterparts, whereas the models manage better their voiceless feature, with confusions leaning towards other voiceless plosives. Overall for the plosives the best fit is provided by the PAR-BLEND model. Likewise, to a large extent listeners' confusion patterns were similar to those of several models for the nasals /n, ɲ/ and the tap /ɾ/, which was frequently confused with /l/.

Nevertheless, some notable differences between listeners and many of the models are visible at the post-test stage. Listeners were highly accurate in identifying the fricative /f/ but all models reported /θ/ for many tokens. On the other hand, listeners confused /θ/ with /s/ which, apart from PAR-ADAPT, the models never did. All models with L1 influence (i.e. apart from SEPARATE) reported /g/ in response to the trill /r/, a confusion never made by listeners. Again, the finding that certain confusions are predicted by different models lends support to the idea that individual L2 sounds may be processed according to different strategies.

5.3. The evolution of individual consonant identification during training

Fig. 7 shows how identification rates change with training day (aggregating training results from each set of 4 sessions) for individual Spanish consonants in listeners and models. Ignoring initial disparities and absolute scores, changes in the pattern of identification rate across training sessions for 14 of the 18 Spanish consonants – viz. /p, t, d, k, g, ʧ, f, θ, x, m, n, ɲ, r, j/ – are well-modelled by one or more strategies, even though these consonants showed a diverse range of individual patterns of improvement or otherwise. For three of the remaining sounds – /b/, /s/ and /l/ – and unlike all the models, listeners barely improved their identification rates with training. For the category /ɾ/, models were unable to mimic the gradual improvement shown by listeners. In some cases, particularly for the plosives /t, d, k/ and the affricate /ʧ/, the SEPARATE model best matched listeners' pattern of improvement over training sessions. For the nasal sounds the closest fit was produced by PAR-ADAPT.

6. Discussion

6.1. L1 influence

Theoretical models such as PAM (Best, 1995), SLM (Flege, 1995) and NLM (Kuhl, 1993) suggest that L1 influences play a critical role in non-native sound perception, a claim that is clearly supported by several outcomes of the current modelling study.

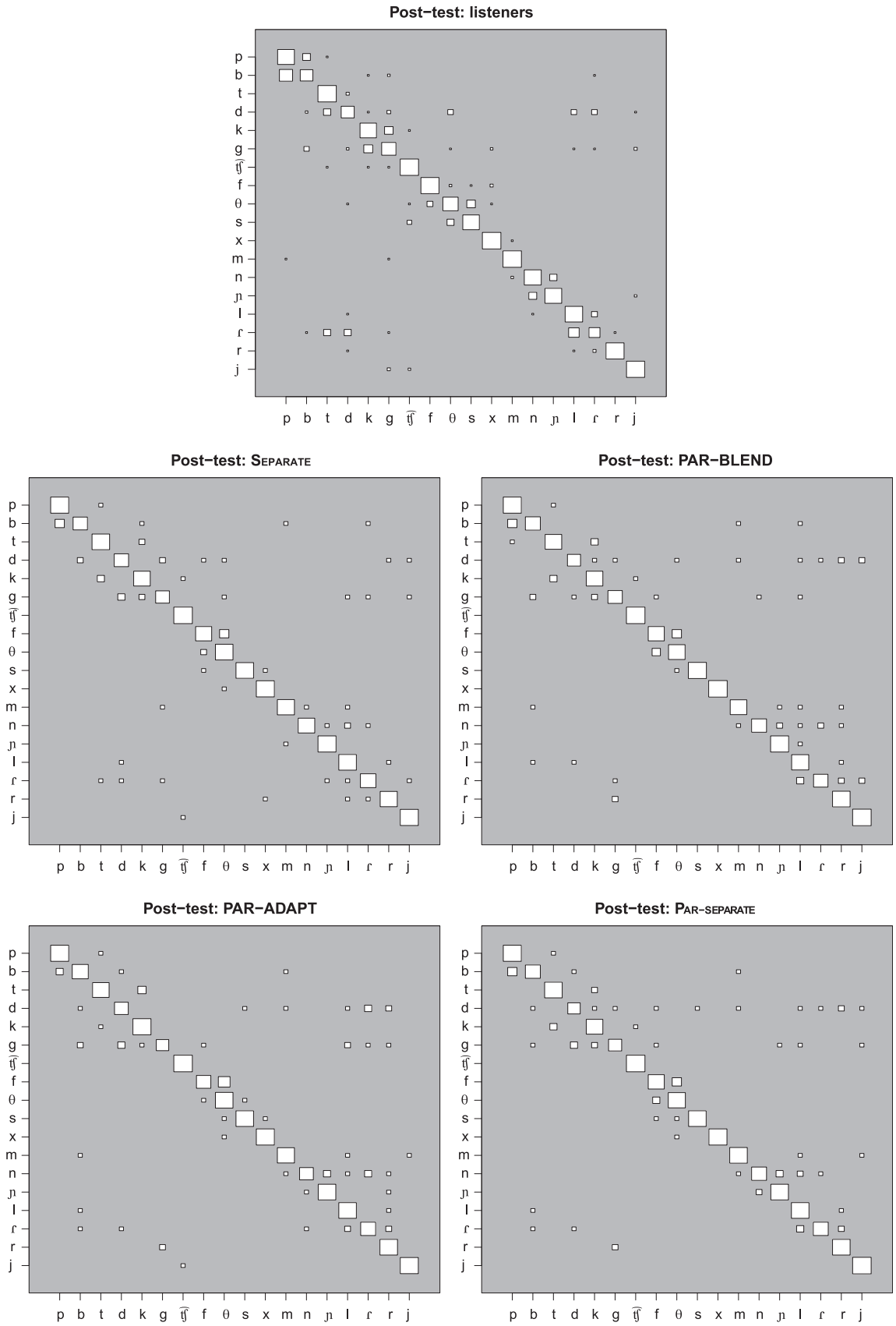

Fig. 6. Consonant confusions in the post-test. Each row depicts responses to the sound labelled on the *y*-axis.

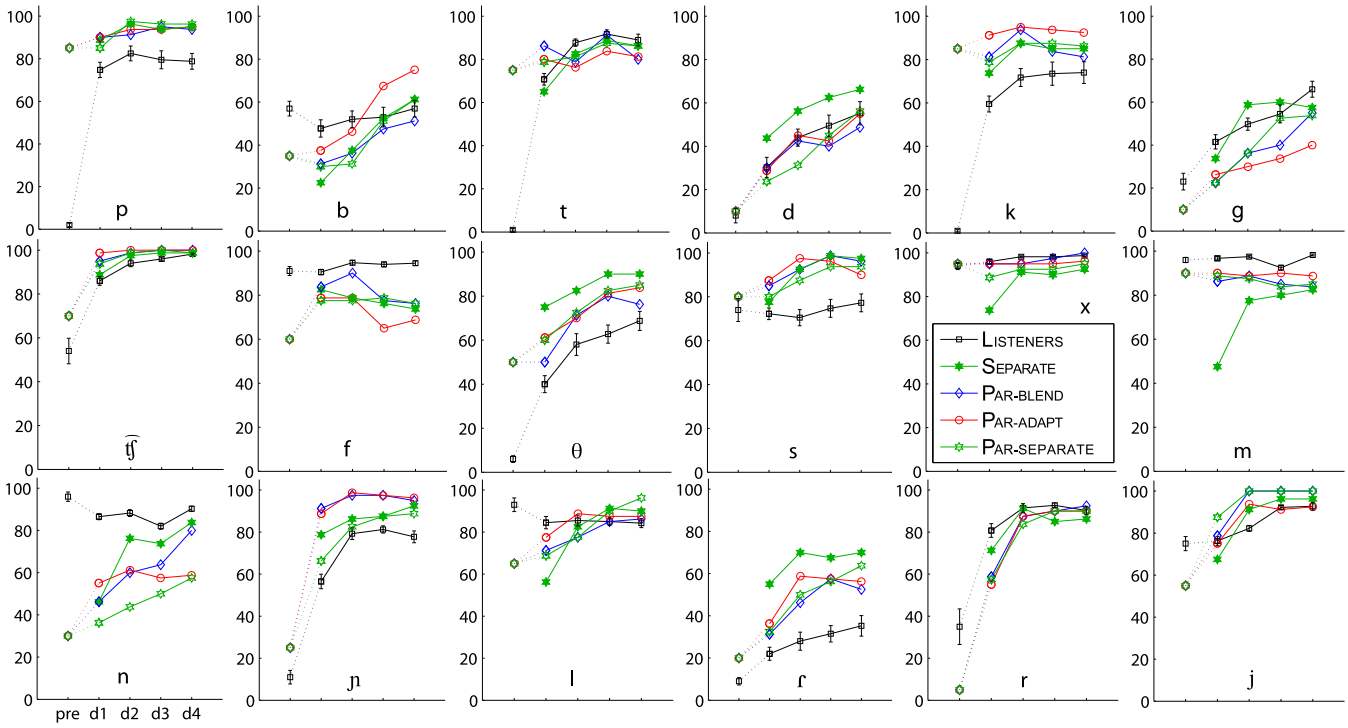

Fig. 7. Model and listener identification rates at each stage of training for individual Spanish consonants. Pre refers to the pre-test scores obtained using the initial models. For clarity, results from the four training sessions on each day are displayed as daily averages (d1, d2, d3, d4).

First, the evolution of consonant identification with supervised exposure (Fig. 4) demonstrates that the SEPARATE model – trained solely on Spanish data – under-predicts human listener scores by around 10 percentage points in the early training stages but then quickly catches up and subsequently over-predicts in the latter stages. This indicates a beneficial effect of prior L1 experience (Flege, 1987, 1995; Iverson et al., 2003; Oh et al., 2011), presumably exploiting acoustic similarities between Chinese and Spanish consonants to identify L2 consonants at levels well above chance. It also suggests a negative subsequent effect of L1 knowledge, perhaps due to the existence of competing similar sounds in the phonological representations of the two languages. This result is consistent with the finding of Aoyama et al. (2004) that adult Japanese speakers exhibit an initial advantage over Japanese children in discriminating English /r/-/l/ and /r/-/w/, but the latter group out-perform adults in subsequent stages of acquisition. A similar initial adult advantage is reported in Oh et al. (2011). Aoyama et al. (2004) suggested that the advantage for adult Japanese listeners might originate in their previous English learning experience, which the Japanese children lacked. It can also be argued that adults' mature L1 phonological systems help them in perceiving sounds that are similar to L1 categories in a non-native language (Oh et al., 2011), since children – with their less mature but more flexible phonological systems – have been found to be less likely than adults to assimilate non-native sounds to a single native category (Baker et al., 2008).

A second way in which L1 influence is apparent comes from differences in the use of training data in the modelling strategies. In the BLEND strategy, models are trained on all the available data (both Chinese and Spanish) at each stage, allowing the model to retain all the exposure from previous learning stages, resulting in an L1 influence that is essentially fixed throughout the training process. The ADAPT strategy, on the other hand, permits new L2 data to incrementally modify the initial pure L1 models. Here, it is possible to argue that L1 influence is more rapidly suppressed under the weight of L2 data. Certainly, the ADAPT model is more prone to over-predicting listener data than the BLEND approach, although the degree of over-prediction reduces with L2 exposure to levels similar to those seen for the BLEND method, suggesting that both techniques end up with models that have a similar degree of L1 influence.

Third, permitting the initial L1-dominated models to operate in parallel with the evolving L2 models in all cases led to better predictions of listener identification scores. The largest improvements were for the PAR-SEPARATE model, followed by the PAR-ADAPT strategy, and with relatively little effect for the PAR-BLEND case. This ranking of gains with the use of parallel models is consistent with the importance of L1 influence, since those models with least L1 influence led to the largest gains when the initial models were available. Further, the generally better fit provided by the parallel models over their non-parallel counterparts suggests that learners' L1 and L2 sound systems are activated simultaneously and cooperate together to influence

perceptual decisions during L2 perception. In fact, evidence exists that a bilingual's L1 and L2 systems are activated at the same time during low level (e.g., phonetic) language processing (Marian et al., 2003). Theoretical models such as SLM postulate that a learner's L1 and L2 categories are situated in a common space and interact with each other constantly (Flege, 1995; Flege et al., 2003), consistent with our simulation results.

6.2. Rapid learning

A key finding of the current study is the ability to simulate listeners' rapid learning, especially for the ADAPT modelling strategy. Previous studies have demonstrated that listeners can modify their existing L1 categories to adapt to ambiguous speech after a short period of exposure (Norris et al., 2003). Rapid adaptations to accented speech or different speakers have also been reported in studies such as Clarke and Garrett (2004), Bradlow and Bent (2008) and Dahan et al. (2008). In fact, rapid learning of non-native sound categories is a feature of many phonetic auditory training studies (Logan et al., 1991; Wang et al., 1999; Lambacher et al., 2005; Nishi and Kewley-Port, 2007).

6.3. Consonant- and listener-specific L2 acquisition strategies

The modelling outcomes of the current study provide clear support for the idea that the manner in which listeners make use of L2 speech input for their evolving sound system is determined by the specific relationship between each L2 sound and the L1 sound system, as is implied by the dominant L2 speech acquisition models mentioned at the start of this section. Different modelling strategies in many cases resulted in distinct predictions of the pattern of consonant identification as a function of training. We give two contrasting examples to illustrate this point. Consistent with SLM, Chinese listeners are likely to create a new category for the Spanish trill /r/, being dissimilar to all Chinese consonants and also not confusable with other Spanish sounds. As a consequence, all the models make similar predictions, suggesting a small L1 influence. However, Spanish /θ/ – even though it too could be seen as a new sound for Chinese listeners – is relatively close acoustically to Chinese /f/ and, in terms of formant transitions, to /s/ (Johnson, 2003). In this case, the creation of a new category (simulated by the SEPARATE model) would predict identification rates higher than those actually observed. Instead, incorporating L1 influence via the parallel models produces a closer match to listeners' identification scores throughout training.

Different individuals may also use differing strategies when faced with non-native sounds, or when placed in different contexts. The alternative models described here might then reflect different strategies and environments. For example, the parallel versions of the models assume an intact L1 system and might be considered to mirror adult L2 learning in a foreign language context while those involving some form of blending of data may be applicable in immersion settings. Likewise, ADAPT seems more representative of learning with a pre-established L1 system while BLEND reflects a situation where more than one language is present, and in differing degrees, in the environment during early phases of learning.

6.4. Limitations

The current study demonstrates what can be achieved using a powerful statistical learning framework to exploit precisely the same L2 data as that heard by listeners. Nevertheless, the modelling approach has a number of limitations. First, the representation of speech in terms of mel-frequency cepstral coefficients and their first and second derivatives, while to some extent validated for automatic speech recognition, differs from that believed to underlie human perception. The use of speech parameters derived from models of the auditory system remains a challenging task, not least due to the need for a 'back-end' recognition architecture well-matched to auditory forms of representation.

Second, the current study is limited to an analysis of the macroscopic features present in model responses, examining the nature of models' improvement in consonant identification over training sessions. Further predictions are possible at the level of individual consonant confusions. However, there is clear evidence of orthographic influence at this level, and extensions to the model to cater for the influence of prior experience with the written form are required before an adequate account of individual confusions can be constructed.

The study also exposes the need for initial models with a closer match to listeners' responses, a deficit that can be interpreted as a lack of good assimilation models.

Finally, the data and results we present are specific to the language pairing (L1: Mandarin; L2: Castillian Spanish) and to consonant learning in a specific context, namely intensive *ab initio* training. Nevertheless, we believe that the modelling approach provides a methodological framework for evaluating other pairings and learning contexts, and further permits the assessment of future progress in overcoming limitations of existing models.

7. Conclusions

First language influence and second language input play a dominant role in theoretical models of how listeners process second language sounds. The current study used computational modelling techniques to explore different assumptions about the detailed processes involved in the acquisition of intervocalic consonants in an unfamiliar language. The outcomes of the study suggest that simulations, when used alongside tightly-controlled listener studies using identical inputs, can be used to test theoretical mod-

els of the role of first language phonology in second language acquisition and additionally to make quantitative predictions at the level of individual sounds of how well alternative models account for behavioural observations.

Acknowledgements

The study was supported by the Basque Government Grant *Language and Speech* (IT311-10) and Spanish Government Grant *DIACEX* FFI2012-31597.

References

Aoyama, K., Flege, J.E., Guion, S.G., Akahane-Yamada, R., Yamada, T., 2004. Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. J. Phonetics 32, 233–250.

Baker, W., Trofimovich, P., Flege, J.E., Mack, M., Halter, R., 2008. Child–adult differences in second-language phonological learning: the role of cross-language similarity. Lang. Speech 51 (4), 317–342.

Best, C.T., 1995. A direct realist view of cross-language speech perception. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues. York Press, Baltimore, pp. 171–204.

Best, C.T., Tyler, M.D., 2007. Nonnative and second-language speech perception: commonalities and complementarities. In: Munro, M., Bohn, O.-S. (Eds.), Language Experience in Second Language Speech Learning: In Honor of James Flege. John Benjamins, Amsterdam, pp. 13–34.

Bradlow, A.R., Bent, T., 2008. Perceptual adaptation to non-native speech. Cognition 106 (2), 707–729.

Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R., Tohkura, Y., 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. J. Acoust. Soc. Am. 101 (4), 2299–2310.

Bundgaard-Nielsen, R.L., Best, C.T., Tyler, M.D., 2011. Vocabulary size matters: the assimilation of second-language Australian English vowels to first-language Japanese vowel categories. Appl. Psycholinguistics 32 (1), 51–67.

Chen, Y., 2007. A comparison of Spanish produced by Chinese L2 learners and native speakers: an acoustic phonetics approach (PhD Dissertation). University of Illinois at Urbana-Champaign.

Clarke, C.M., Garrett, M., 2004. Rapid adaptation to foreign-accented English. J. Acoust. Soc. Am. 116 (6), 3647–3658.

Dahan, D., Drucker, S.J., Scarborough, R.A., 2008. Talker adaptation in speech perception: adjusting the signal or the representations?. Cognition 108 (3) 710–718.

de Boer, B., Kuhl, P.K., 2003. Investigating the role of infant-directed speech with a computer model. Acoust. Res. Lett. Online 4 (4), 129–134.

Díaz-Campos, M., 2004. Context of learning in the acquisition of Spanish second language phonology. Stud. Sec. Lang. Acquis. 26, 249–273.

Díaz-Campos, M., 2014. Segmental phonology in second language Spanish. In: Geeslin, K. (Ed.), The Handbook of Spanish Second Language Acquisition. Wiley& Sons, Chichester, pp. 146–165.

Escudero, E., Kastelein, J., Weiand, K., van Son, R., 2007. Formal modelling of L1 and L2 perceptual learning: computational linguistics versus machine learning. In: Proceedings of Interspeech 2007, pp. 1889–1892.

Flege, J.E., 1987. The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. J. Phonetics 15 (1), 47–65.

Flege, J.E., 1995. Second-language speech learning: theory, findings, and problems. In: Strange, W. (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research. York Press, Baltimore, pp. 233–277.

Flege, J.E., Schirru, C., MacKay, I.R.A., 2003. Interaction between the native and second language phonetic subsystems. Speech Commun. 40 (4), 467–491.

Gauvain, J.-L., Lee, C.-H., 1991. Bayesian learning of Gaussian mixture densities for hidden Markov models. In: Proc. DARPA Speech Natural Language Workshop (Pacific Grove), pp. 272–277.

Gauvain, J.-L., Lee, C.-H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Process. 2 (2), 291–298.

Gil Fernández, J., 2007. Fonética para profesores de espa nol: de la teoría a la práctica. Madrid, Arco/Libros.

Gong, J., 2013. Computational modelling of sound perception in a second language (PhD Dissertation). University of the Basque Country. <http://www.laslab.org/resources/Thesis_JianGong.pdf>.

Gong, J., Cooke, M., García Lecumberri, M.L., 2010. Towards a quantitative model of Mandarin Chinese perception of English consonants. In: Dziubalska-Kolaczyk, K., Wrembel, M., Kul, M. (Eds.), New Sounds 2010: Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech. Adam Mickiewicz University, Poznan,Poland, pp. 580–584.

Gong, J., Cooke, M., García Lecumberri, M.L., 2011. A computational modelling approach to the development of L2 sound acquisition. In: Proceedings of the ICPhS 2011, pp. 755–758.

Hualde, J.I., 2005. The Sounds of Spanish. Cambridge University Press, Cambridge.

Iverson, P., Kuhl, P.K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., Siebert, C., 2003. A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition 87, B47–B57.

Johnson, K., 2003. Acoustic and Auditory Phonetics, 2nd ed. Blackwell.

Kuhl, P.K., 1993. Innate predispositions and the effects of experience in speech perception: the native language magnet theory. In: de Boysson-Bardies, B., de Schonen, S., Jusczyk, P., MacNeilage, P., Morton, J. (Eds.), Developmental Neurocognition: Speech and Face Processing in the First Year of Life. Springer, Berlin, pp. 259–274.

Lake, B.M., Vallabha, G.K., McClelland, J.L., 2009. Modeling unsupervised perceptual category learning. IEEE Trans. Auton. Ment. Dev. 1 (1), 35–43.

Lambacher, S.G., Martens, W.L., Kakehi, K., Marasinghe, C.A., Molholt, G., 2005. The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. Appl. Psycholinguistics 26 (02), 227–247.

Lively, S.E., Logan, J.S., Pisoni, D.B., 1993. Training Japanese listeners to identify English /r/ and /l/. II: the role of phonetic environment and talker variability in learning new perceptual categories. J. Acoust. Soc. Am. 94 (3), 1242–1255.

Logan, J.S., Lively, S.E., Pisoni, D.B., 1991. Training Japanese listeners to identify English /r/ and /l/: a first report. J. Acoust. Soc. Am. 89 (2), 874–886.

Marian, V., Spivey, M., Hirsch, J., 2003. Shared and separate systems in bilingual language processing: converging evidence from eyetracking and brain imaging. Brain Lang. 86 (1), 70–82.

Martínez-Celdrán, E., Fernández-Planas, A.M., Carrera-Sabaté, J., 2003. Castilian Spanish. J. Int. Phonetic Assoc 33 (02), 255–259.

McMurray, B., Aslin, R.N., Toscano, J.C., 2009. Statistical learning of phonetic categories: insights from a computational approach. Dev. Sci. 12 (3), 369–378.

Morrison, G.S., 2009. L1-Spanish Speakers' Acquisition of the English /i/-/ɪ/ Contrast II: Perception of Vowel Inherent Spectral Change. Lang. Speech 52 (4), 437–462.

Nishi, K., Kewley-Port, D., 2007. Training Japanese listeners to perceive American English vowels: influence of training sets. J. Speech Lang. Hearing Res. 50 (6), 1496–1509.

Norris, D., McQueen, J.M., Cutler, A., 2003. Perceptual learning in speech. Cogn. Psychol. 47 (2), 204–238.

Oh, G., Guion-Anderson, S., Aoyama, K., Flege, J., Akahane-Yamada, R., Yamada, T., 2011. A one-year longitudinal study of English and

Japanese vowel production by Japanese adults and children in an English-speaking setting. J. Phonetics 39 (2), 156–157.

Quilis, A., 1997. Principios de fonología y fonética espa nolas. vol. 43. Madrid, Arco/Libros.

Strange, W., Bohn, O.-S., Trent, S.A., Nishi, K., 2004. Acoustic and perceptual similarity of North German and American English vowels. J. Acoust. Soc. Am. 115 (4), 1791–1807.

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

Vallabha, G.K., McClelland, J.L., Pons, F., Werker, J.F., Amano, S., 2007. Unsupervised learning of vowel categories from infant-directed speech. Proc. Nat. Acad. Sci. 104 (33), 13273–13278.

Wang, Y., Spence, M.M., Jongman, A., Sereno, J.A., 1999. Training American listeners to perceive Mandarin tones. J. Acoust. Soc. Am. 106 (6), 3649–3658.

Young, S.J., Evermann, G., Gales, M., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book Version 3.4. Cambridge University Engineering Department.

Zampini, M.L., 1994. The role of native language transfer and task formality in the acquisition of Spanish spirantization. Hispania 77, 470–481.

Zampini, M.L., 2014. Voice onset time in second language Spanish. In: Geeslin, K. (Ed.), The Handbook of Spanish Second Language Acquisition. Wiley& Sons, Chichester, pp. 111–129.