ASR Systems as Models of Phonetic Category Perception in Adults

Thomas Schatz[1,2] **(thomas.schatz@laposte.net)**
Francis Bach[2] **(francis.bach@ens.fr)**
Emmanuel Dupoux[1] **(emmanuel.dupoux@gmail.com)**
[1]LSCP (ENS/EHESS/CNRS), DEC, ENS, PSL Research University, France
[2]SIERRA Project-Team (ENS/INRIA/CNRS), DI, ENS, PSL Research University, France

Abstract

We test the potential of standard Automatic Speech Recognition (ASR) systems trained on large corpora of continuous speech as quantitative models of human speech processing. In human adults, speech perception is attuned to efficiently process native speech sounds, at the expense of difficulties in processing non-native sounds. We use ABX-discriminability measures to test whether ASR models can account for the patterns of confusion between speech sounds observed in humans. We show that ASR models reproduce some well-documented effects in non-native phonetic perception. Beyond the immediate results, our methodology opens up the possibility of a more systematic investigation of phonetic category perception in humans.

Keywords: non-native speech perception, perceptual atunement, psycholinguistics, computational modeling

Introduction

In this paper, we ask to what extent ASR systems, considered as computational models of speech processing in human adults, can account for the well-documented influence of the native language on foreign language phonetic category perception how phonetic categories from foreign languages are perceived (for instance Strange, 1995 or Cutler, 2012). For example, native speakers of Japanese, a language where the /r/ and /l/ sounds from American English are not contrastive, have a very hard time discriminating between these two sounds (Goto, 1971; Miyawaki et al., 1975). We restrict ourselves to the question of the perception of foreign phonetic categories by monolingual listeners, who have no or very little prior experience with speech in other languages. In particular, we do not discuss the question of native language (L1) influence on phonetic categories perception during second language (L2) learning.

We obtain a computational model of speech processing by speakers of language *A*, by training an ASR system with annotated speech in that language. This model is then presented with audio recordings in another language *B*, and the discriminability between phonetic categories in this other language is evaluated on the resulting output using ABX-discriminability measures (Schatz et al., 2013; Schatz, 2016). This allows us to evaluate the model by comparing the patterns of discriminability it predicted with available empirical evidence regarding the confusions made by native speakers of language *A* between the phonetic categories of language *B*. We study the pattern of cross-linguistic phonetic category confusions in four languages : American English, Japanese, Mandarin and Vietnamese. Given the space constraints in this paper we have to restrict ourselves to showcase a limited but representative sample of the kind of analyses that can be carried out with our methodology. See Schatz (2016) for more.

Existing theoretical frameworks for cross-linguistic phonetic category perception such as the Perceptual Assimilation Model (PAM) (Best, 1995) the Speech Learning Model (SLM) (Flege, 1995) and the Native Language Magnet model (NLM) (Kuhl & Iverson, 1995) rely on an unspecified mapping of foreign phonetic categories onto native phonetic categories to make predictions about the discriminability of phonetic contrasts. In this sense, they only specify a *discrimination task model*, that can then be applied to different *speech processing models*, where a *speech processing model* specifies a particular way of mapping foreign phonetic categories onto native phonetic categories. Conceptual speech processing models based on an analysis of the phonology of the languages involved have been tested, with disappointing results (Strange, Bohn, Trent, & Nishi, 2004). In particular, an abstract analysis at the level of phonemes appears doomed to fail, because phonetic and acoustic details in the stimuli, for example related to the phonetic and prosodic context, have been observed to influence the discriminability of segments. This leads to the idea of using models based on the processing of actual speech recordings. We found only two previous studies using such models to look at cross-linguistic phonetic category perception by monolingual speakers: Strange et al. (2004) and Gong, Cooke, and Garcia Lecumberri (2010). Strange and colleagues tried to predict cross-linguistic assimilation patterns of North German vowels into American English vowels by monolingual speakers of American English. Gong and colleagues tried to predict cross-linguistic assimilation patterns of English consonants into Mandarin consonants by monolingual speakers of Mandarin. The two studies report contrasting results. Gong et al. found a good match between their model's predictions and empirical observations, while Strange et al. found many discrepancies.

There are three main differences between the research presented in this paper and previous approaches. First, we replace discriminability predictions based on the PAM framework with ABX-discriminability measures. Second, we replace ad hoc speech processing models trained on very specific stimuli with general purpose ASR systems trained on natural continuous speech. Third, we look at the full inventory of phonetic categories for four different languages. Let us motivate each of this points separately.

Using ABX-discriminability measures has two main benefits. First, ABX discriminability measures can be directly

related to the ABX discrimination tasks commonly used to study cross-linguistic phonetic category perception (Schatz, 2016). Second, ABX discriminability measures can be seen as a generalization of predictions derived from PAM in the sense that, given a representation of speech segments consisting of a category label together with a category goodness rating, it is easy to provide a dissimilarity function such that PAM-based predictions and ABX discriminability measures computed with this dissimilarity function are compatible. The interest of the generalization is that unlike PAM-based predictions, that require categorical representations with category goodness ratings, ABX discriminability measures can be applied to any kind of representation for which a dissimilarity function can be provided.

Using general-purpose ASR systems has also several important benefits. First, it seems more natural to use as a model of speech processing in humans a general purpose system rather than a system only trained on isolated VCV stimuli with limited vocalic variability as in Gong et al. (2010) or a system that can only recognize vowels and which uses speech features (F1/F2/F3 formants and duration) and an acoustic model (LDA) that are known not to be very performant for speech recognition purposes as in Strange et al. (2004). Second, the ability of the system to handle natural continuous speech, means that it can capitalize on the many existing corpora of annotated speech recordings. This allows training and testing systems in many languages and in a much more extensive manner. Third, using ASR systems means that the results are also of interest for the ASR community. Indeed, discrepancies found between ASR systems predictions and human behavior can provide insight into the shortcomings of ASR systems and inspiration for improving them.

By training and testing systems in four languages, we are able to evaluate the models in a much more comprehensive way. We are able to investigate both global effects in the perception of phonetic categories (for example phonetic contrasts of a language are globally harder to discriminate for non-native speakers than for native speakers (Gottfried, 1984)) and more local effects (for example related to the perception of American English /r/-/l/ by Japanese listeners (Goto, 1971; Miyawaki et al., 1975))

Methods

Corpora

To train and evaluate ASR models, 5 corpora of recorded speech in different languages were used: a subset of the Wall Street Journal corpus (WSJ) (Paul & Baker, 1992), the Buckeye corpus (BUC) (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005), a subset of the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003), the Global Phone Mandarin (GPM) corpus (Schultz, 2002) and the Global Phone Vietnamese (GPV) corpus (Vu & Schultz, 2009). Important characteristics of the various corpora are summarized in Table 1. Two corpora in American English were included to allow the separation of effects due to a change in language from effects due to other kinds of difference between corpora (e.g. the properties of the recording microphones, speech register, etc.). Phonetic transcriptions were obtained from phonetic dictionaries and word-level transcriptions for the WSJ, BUC, GPM and GPV corpora. For the CSJ corpus, manual phonetic transcriptions were used. For all corpora, timestamps for the phonetic transcriptions were obtained by forced-alignment using a GMM-HMM ASR system similar to those described in the next section, but trained on the whole corpus.

ASR Models

We restrict our investigation to Gaussian-Mixture based Hidden-Markov-Model (GMM-HMM) architectures by opposition to Neural-Network based architectures. Each corpus was randomly split into a training and a test set, each containing an equal number of speakers. Then, a GMM-HMM ASR model was trained with the Kaldi toolkit (Povey et al., 2011) on the training set of each corpora. The same Kaldi recipe was used (see `https://github.com/bootphon/abkhazia/blob/master/abkhazia/kaldi/kaldi_templates/train_and_decode.sh`) with the same parameters and input features to train all models. Input features consisted of 13 MFCC coefficients plus 3 pitch-related features (Ghahremani et al., 2014) and their delta and delta-deltas coefficients. Pitch features were included because tone is contrastive in Mandarin and Vietnamese (i.e. there are words differing only by their tone in these languages). The Word Error Rate (WER) on the test set for each of the resulting modelsis reported in Table 1. The best performance is obtained on the WSJ corpus and the worse on the BUC corpus. The bad performance on the BUC corpus is probably due to its particularly casual register.

Table 1: Word-Error-Rates obtained by the ASR systems trained on each corpus as well as the language, total duration, speech register and number of speakers for each corpus. AE stands for American English, Spont. stands for Spontaneous.

Corpus	Language	Time	Type	Spk	WER
WSJ	AE	143h	Read	338	8.5%
BUC	AE	19h	Spont.	40	48.0%
CSJ	Japanese	15h	Spont.	75	30.0%
GPM	Mandarin	30h	Read	132	31.0%
GPV	Vietnamese	20h	Read	129	23.5%

GMM-HMM models can provide an output in many different formats, such as word- or phone-level transcriptions, word- or phone-level *n*-best lists or lattices or frame-by-frame phone-level posteriorgrams of various kinds. We consider only Viterbi-smoothed phone-level posteriorgrams, which are more informative than phone-level transcriptions. They are obtained using a phone-level bigram language model estimated on the training set of each corpus.

ABX Evaluation

The test set of each corpus is decoded with each of the 5 ASR models, producing Viterbi posteriorgrams, and the input features are added as a control, yielding a total of 6 different representations for each of the 5 corpora. For each corpus, a minimal-pair ABX task (Schatz, 2016; Schatz et al., 2013) ON phonetic category BY talker, previous phone and following phone is compiled on the test set and used to measure the ABX-discriminability between phonetic categories for the 6 representations of that corpus. The basic idea behind this evaluation method is to form triplets A, B, X of stimuli such that X is supposed to be more like A than like B based on external labels and to look whether it is indeed the case that the representations a, b, x of A, B and X by the model to be evaluated are such that $d(a,x) < d(b,x)$ for some measure of dissimilarity d on the space of representations. In the specific ABX task we consider here, the stimuli are phone occurrences in a given test set and ABX triplets are formed such that A and B are occurrences of different phones, X is an occurrence of the same phone as A and the preceding phone, following phone and talker for A, B, and X are the same. A triplet in such a task could be for example:

A	B	X
/i/$^{T_1}_{b\_t}$	/u/$^{T_1}_{b\_t}$	/i/$^{T_1}_{b\_t}$

where $_{b\_t}$ indicates a segment preceded by a /b/ and followed by /t/ and ^{T_1} indicates a segment pronounced by speaker T_1. The ABX-discriminability error rate for a given choice of the A,X phone, the B phone, the preceding and following phones and the talker is obtained as the proportion of corresponding triplets for which $d(a,x) > d(b,x)$. A summary ABX error rate for a given phonetic contrast, for example /u/-/i/, is obtained by averaging first over the error rates obtained when A and X are chosen to be /u/ and B is chosen to be /i/ and vice-versa, then averaging over all possible choice of talker and finally averaging over all possible choices of preceding and following phones. Depending on the experiments, we use directly the scores obtained for individual phonetic contrasts or we average them over interesting classes of contrasts (e.g. all vowel contrasts)

Dissimilarities between the sequences of posteriorgrams corresponding to phones are computed using Dynamic Time Warping (DTW) (Vintsyuk, 1968) based on a frame-to-frame symmetric Kullback-Leibler divergence (Kullback & Leibler, 1951). Note that, unlike what is done in PAM, dissimilarities are here based on more than just category labels or category-goodness of the most likely categories. For the input representation control, we compute dissimilarities using DTW based on a frame-to-frame cosine distance. We do not claim that these choices of dissimilarity functions are necessarily optimal and it would be interesting to test other possibilities.

Analyses

We test whether ASR models can account for various effects in cross-linguistic perception of phonetic categories. Our methodology allows us to study both global effects, that involve all possible phonetic contrasts and allow for a systematic evaluation of the models, and more local effects, that involve specific phonetic contrasts and for which extensive empirical data is often available. Because of space constraints, we restrict our presentation to a limited but representative sample of possible analyses. See Schatz (2016) for more.

As a first global measure, we compute the average ABX error rates over all consonant contrasts and over all vowel contrasts in the corpus. Phonetic contrasts of a language are found to be globally harder to discriminate for non-native listeners than for native listeners (see e.g. Gottfried (1984)). We should thus find that the phonetic contrasts of a language are globally harder to discriminate for models trained on a different language (mismatched-language condition) than on the same language (matched-language condition).

Although phonetic contrasts are globally harder to discriminate for non-native listeners than for native listeners, it is well-established that all contrasts are not affected to the same extent. The way in which the contrasts are differentially affected is known to be largely determined by the native language of the non-native listeners (Strange, 1995). In our setting, this predicts that the patterns of confusion for the representations obtained from the two American English ASR models should be more similar to each other than to representations obtained from other ASR models, and this independently of the corpus on which the confusions are measured. To test this, we devise a second global measure. For a given model m and a given test corpus c, we look at the vector $v_{m,c}$ composed of the ABX error rates obtained with this model on each of the vowel contrasts present in the test corpus. For each pair of models m_1, m_2 and each corpus c, we then compute $s_c(m_1,m_2)$ the cosine similarity between patterns of confusions $v_{m_1,c}$ and $v_{m_2,c}$. By averaging $s_c(m_1,m_2)$ over our five different corpora, we obtain a global measure of the similarity between the pattern of confusions between phonetic categories predicted by the models m_1 and m_2. We also define a similar measure based on consonant contrasts instead of vowel contrasts. Note that, thanks to the scale invariance of the cosine similarity, this second global measure compares pattern of errors independently of the average levels of errors involved, i.e. there is zero redundancy with the first global measure we proposed. The results obtained with the second measure take the form of similarity matrices between models. To visualize these results conveniently, we plot 2D embeddings of the models obtained through Multi-Dimensional Scaling (MDS) applied to the similarity matrices. The embeddings are obtained with the scikit-learn (Pedregosa et al., 2011) implementation of non-metric MDS.

Finally, we also investigate a more local effect. The perception of American English /r/-/l/ by Japanese listeners is known to be very poor (Goto, 1971; Miyawaki et al., 1975); we therefore expect a model trained on Japanese stimuli (CSJ model) to be worse at discriminating this contrast than any model trained on American English (WSJ and BUC models).

Results

Global Effects

Average ABX-discriminability for American English consonant contrasts are shown on Figure 1. Results for American English vowels, and both vowels and consonants in other languages are not shown here, but are fully consistent with our discussion (see Schatz (2016)). We see that the model performing best on a given corpus is always the model that was trained on (independent) stimuli from that same corpus. We also see that the two models trained on a corpus of American English (WSJ and BUC) separate phonetic categories in that language better than phonetic categories from other languages, even when they are tested on a different corpus than the one on which they were trained. This means that the differences observed cannot be explained by low-level differences in the corpora, such as the type of microphone used to record the signal for example. This is all the more interesting since the WSJ corpus is actually more similar in many respects to the other corpora, in particular the GPM and GPV corpora, than to the BUC corpus. For example, the speech register and the topics are similar in the WSJ, GPM and GPV corpora (news articles readings) and different from the speech register and topics in the BUC corpus (spontaneous, and often quite casual, dialog with an interviewer on everyday topics) and in the CSJ corpus (relations, not read but somewhat prepared, of a memorable episode of their life by speakers in front of a small audience). The results cannot be explained either by an overall better quality of the two English models independently of language, since, as we saw in Table 1, the BUC model has the worse performance of all models in terms of WER. This overall suggests that the ASR models succeeded in learning a truly *language-specific* representation. Another interesting observation is that training a model on a different language appears to render phonetic categories less separable than they were in the input representation. A simple interpretation is that the categorical representations that are learnt by the ASR systems have the benefits of providing much more compressed representation of the speech signal than the input features, at the expense of a loss of representation power for stimuli in languages other than the one used for training the system.

Next, we ask whether the difference in average discriminability between models in matched-language and mismatched-language conditions results from a global rescaling of the discriminability of each phonetic contrasts or whether, as expected from phonetic category perception in human, certain contrasts are more impacted than others depending on the peculiarities of the training and test languages. Each model is represented in Figure 2 as a point in a two-dimensional embedding obtained from the similarities between the pattern of errors of the different models for consonant contrasts (left) or vowel contrasts (right) as described in the Methods. The most interesting result is that for both consonants and vowels, the error patterns of the BUC and WSJ models are more similar to each other than to any other

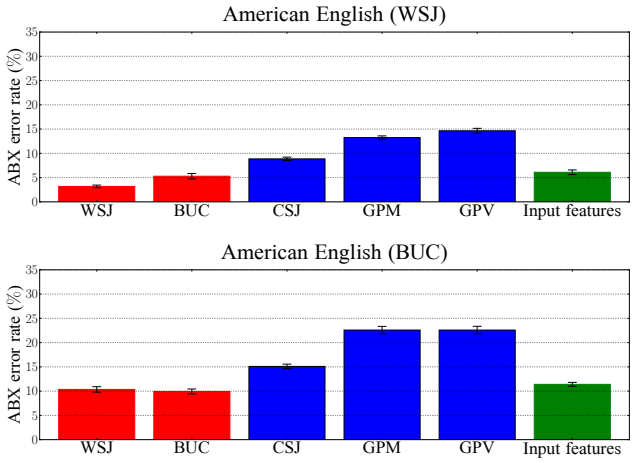

Figure 1: Average ABX error-rates for all consonant contrasts of American English. The first line corresponds to models tested on stimuli from the WSJ corpus, the second line to models tested on stimuli from the BUC corpus. Red bars indicate *matching language* conditions, i.e. where a model's *native language* (i.e. the language in the corpus on which the model was trained) matches the language in the test corpus. Blue bars indicate *mismatched language* conditions. Green bars are used for the baseline obtained by taking as a representation the input features common to all models (i.e. MFCC plus pitch features), without any language-specific processing. Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores over the different speakers in each corpus.

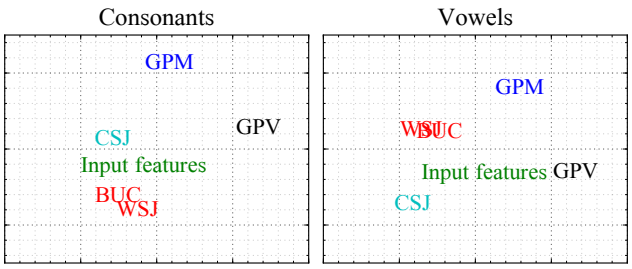

Figure 2: Left: two-dimensional embeddings of the different models based on the cosine similarity between their patterns of errors for consonant contrasts across the five test corpora (See Methods for details). Right: same for vowel contrasts across the five test corpora.

model (and no other pair of models are closer than these two). This indicates that the patterns of errors obtained are *language-specific*. Another result is that the baseline error patterns obtained with language-independent input features appear rather central, although they tend to be closer to the error patterns obtained with the WSJ, BUC and CSJ models than to those obtained with the GPM and GPV models.

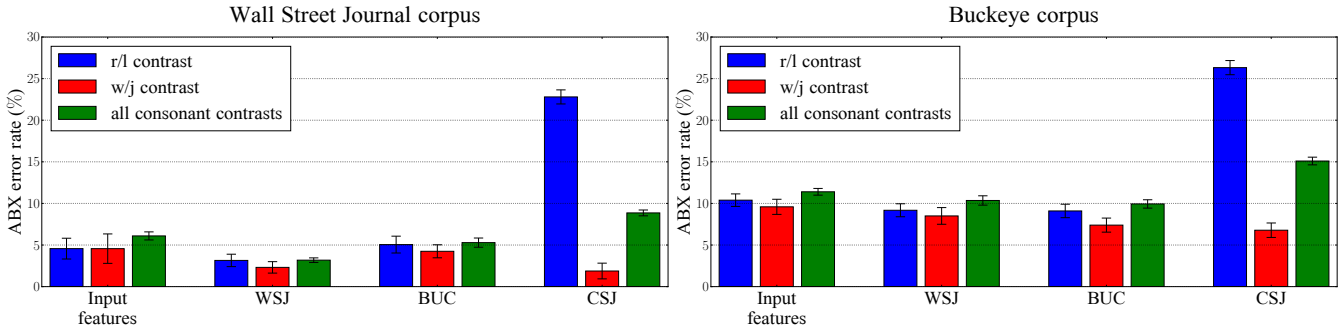

Figure 3: ABX error-rates obtained for the American English /r/-/l/ contrast and two controls using stimuli from the WSJ corpus (left) or from the BUC corpus (right). Four models are tested: language-neutral input features (MFCC+pitch), two *American English* models (WSJ and BUC) and a *Japanese* model (CSJ). Error bars indicates mean plus and minus one standard deviation and were obtained by bootstrap resampling of the scores over the different speakers in each corpus.

Local Effects

To further investigate the potential of these models to account for phonetic category perception, we now look at a more specific effect that has been studied directly in humans. We look at the discriminability of the /r/-/l/ contrast of American English by models trained on American English and on Japanese. This distinction should be much harder to make for the Japanese-trained model than for models trained on American English (Goto, 1971; Miyawaki et al., 1975). In Figure 3, we plotted the ABX error-rates for the /r/-/l/ contrast obtained on the WSJ corpus and on the BUC corpus. We also plotted as controls the error-rate for /w/-/j/, another liquid contrast, and the average error-rate for all consonant contrasts. We see that the discriminability of phonetic categories is globally more difficult in the BUC corpus than in the WSJ corpus, but that the same pattern is observed for both corpora, confirming that we are observing language-specific effects and not corpus-specific effects. Looking at the input features baseline, we see that both liquid contrasts are slightly above the average discriminability of consonant contrasts with the /w/-/j/ contrast perhaps slightly easier than the /r/-/l/ contrast. When we look at the matching-language conditions, we see that in all cases the discriminability either improves or at least remains similar when compared to the input features discriminability. In the mismatched-language condition (i.e. for the CSJ model), we see that while the /w/-/j/ contrast becomes even more discriminable than for models trained on American English, the /r/-/l/ contrast becomes much much more difficult to discriminate. The extent of the degradation in the discriminability of the /r/-/l/ contrast is underlined by the comparing it with the degradation in discriminability averaged over all consonants which exists but is much smaller.

Conclusion

In this study, we used ABX discriminability measures to show that standard GMM-HMM ASR systems, viewed as computational models of human speech perception, can account for a variety of empirically observed effects in cross-linguistic phonetic category perception by monolingual speakers. We showed that these systems can account for two types of *global* effects: first, that the phonetic categories of a language are overall harder to discriminate for non-native speakers than for native speakers and second, that the pattern of confusions between phonetic categories for non-native speakers is specific to their native language (e.g. native speakers of Japanese do not make the same confusions between phonetic categories of American English than native speakers of French). We also showed that GMM-HMM ASR systems can account for a well known *local* effect: American English /r/ and /l/ are very hard to discriminate for native speakers of Japanese.

There are several avenues for future work. Global effects could be confirmed using more corpora and a wider range of local perceptual effects could be tested using the same ABX task or variants. For example, a task comparing CC consonant clusters versus CVC triphones of American English could be used to test for the presence of vowel epenthesis in the model trained on a Japanese dataset (Dupoux, Kakehi, Hirose, Pallier, & Mehler, 1999). In addition, exploratory analyses could generate novel predictions. For example, Schatz (2016), showed that GMM-HMM ASR models predict that while native speakers of Vietnamese should be better than native speakers of American English in discriminating the tones of Mandarin, native speakers of Mandarin should be neither better nor worse than native speakers of American English in discriminating the tones of Vietnamese. It would be interesting to test empirically this prediction. Another interest of our method is that it can measure very fine grained phonetic effects, for instance, the difficulty of Japanese native speakers with the /r/-/l/ contrast of American English is known to be stronger in syllable-initial position than in syllable-final position. This could be taken into account by introducing the position in syllable as a BY factor in the design of the ABX task. Going even further, the exact same stimuli could be used to probe models and humans, with similar ABX paradigms. One of the strength of our approach is also that different ASR systems, speech representations, and dissimilarity func-

tions can be tested and quantitatively compared. One outstanding question, for example, is whether Neural Network ASR systems, which have better performance than GMM-HMM ASR systems, are also more accurate models of human speech perception. For instance, GMM-HMM ASR systems are known to be limited in their ability to model phoneme durations accurately (Pylkkönen & Kurimo, 2004). Schatz (2016) showed that they fail to predict that vowel duration contrasts in Japanese are easier to discriminate by speakers of Japanese, than by speakers of American English. It would be interesting to see whether Deep Neural Network (DNN) systems that do not suffer from the same theoretical limitation in modeling phone durations fare better on this specific test. Testing DNN systems would also be interesting from the ASR point of view, since their strengths or weaknesses are less well known that those of the HMM-GMM systems. Finally, because they are trained in a supervised fashion, ASR systems can model phonetic category *perception* in adults, but not phonetic category *acquisition* in infants. It would thus be interesting to test models trained in a more plausible fashion.

Acknowledgments

The research reported here received funding from the European Research Council under the FP/2007-2013 program / ERC Grant Agreement n. ERC-2011-AdG-295810 BOOTPHON and from the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-0001-02 PSL*, ANR-10-LABX-0087 IEC).

References

Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 171–204.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Mit Press.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology-HPP*, *25*(6), 1568-1578.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, 233–277.

Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., & Khudanpur, S. (2014). A pitch extraction algorithm tuned for automatic speech recognition. In *Proc. ICASSP*.

Gong, J., Cooke, M., & Garcia Lecumberri, M. (2010). Towards a quantitative model of mandarin chinese perception of english consonants. *Proc. NewSounds 2010*.

Goto, H. (1971). Auditory perception by normal japanese adults of the sounds l and r. *Neuropsychologia*, *9*(3), 317–323.

Gottfried, T. L. (1984). Effects of consonant context on the perception of french vowels. *Journal of Phonetics*, *12*(2), 91–114.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the perceptual magnet effect. *Speech perception and linguistic experience: Issues in cross-language research*, 121–154.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 79–86.

Maekawa, K. (2003). Corpus of spontaneous japanese: Its design and evaluation. In *Proc. ISCA & IEEE workshop on spontaneous speech processing and recognition.*

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english. *Perception & Psychophysics*, *18*(5), 331–340.

Paul, D. B., & Baker, J. M. (1992). The design for the wall street journal-based csr corpus. In *Proc. workshop on speech and natural language* (pp. 357–362).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. (2005). The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, *45*(1), 89–95.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others (2011). The kaldi speech recognition toolkit. In *Proc. workshop on automatic speech recognition and understanding.*

Pylkkönen, J., & Kurimo, M. (2004). Duration modeling techniques for continuous speech recognition. In *Proc. INTERSPEECH.*

Schatz, T. (2016). *ABX-Discriminability Measures and Applications*. Doctoral dissertation, Université Paris 6 (UPMC).

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proc. INTERSPEECH.*

Schultz, T. (2002). Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Proc. INTERSPEECH.*

Strange, W. (1995). *Speech perception and linguistic experience: Issues in cross-language research*. York Press.

Strange, W., Bohn, O.-S., Trent, S. A., & Nishi, K. (2004). Acoustic and perceptual similarity of north german and american english vowels. *The Journal of the Acoustical Society of America*, *115*(4), 1791–1807.

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, *4*(1), 52–57.

Vu, N. T., & Schultz, T. (2009). Vietnamese large vocabulary continuous speech recognition. In *Proc. ASRU.*