Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception

Thomas Schatz, Francis Bach, and Emmanuel Dupoux

Citation: The Journal of the Acoustical Society of America **143**, EL372 (2018); doi: 10.1121/1.5037615

View online: https://doi.org/10.1121/1.5037615

View Table of Contents: http://asa.scitation.org/toc/jas/143/5

Published by the Acoustical Society of America

Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception

Thomas Schatz[a)]
Department of Linguistics and UMIACS, University of Maryland, College Park, Maryland 20742, USA
thomas.schatz@laposte.net

Francis Bach
SIERRA project-team, Département d'informatique de l'ENS, École normale supérieure, INRIA, CNRS, PSL Research University, 45, rue d'Ulm 75005 Paris, France
francis.bach@ens.fr

Emmanuel Dupoux
LSCP, Département d'études cognitives de l'ENS, École normale supérieure, EHESS, CNRS, PSL Research University, 29, rue d'Ulm 75005 Paris, France
emmanuel.dupoux@gmail.com

Abstract: Theories of cross-linguistic phonetic category perception posit that listeners perceive foreign sounds by mapping them onto their native phonetic categories, but, until now, no way to effectively implement this mapping has been proposed. In this paper, Automatic Speech Recognition systems trained on continuous speech corpora are used to provide a fully specified mapping between foreign sounds and native categories. The authors show how the *machine ABX* evaluation method can be used to compare predictions from the resulting quantitative models with empirically attested effects in human cross-linguistic phonetic category perception.

[DDO]
Date Received: December 7, 2017 **Date Accepted:** April 20, 2018

1. Introduction

The way we perceive phonetic categories (i.e., basic speech sounds such as consonants and vowels) is largely determined by the language(s) to which we were exposed as a child. For example, native speakers of Japanese have a hard time discriminating between American English (AE) /ɹ/ and /l/, a phonetic contrast that has no equivalent in Japanese (Goto, 1971; Miyawaki *et al.*, 1975). Perceptual specialization to the phonological properties of the native language has been extensively investigated using a variety of techniques [see Strange (1995) and Cutler (2012) for reviews]. Many of the proposed theoretical accounts of this phenomenon concur that foreign sounds are not perceived faithfully, but rather, are "mapped" onto one's pre-existing (native) phonetic categories, which act as a kind of "filter" resulting in the degradation of some non-native contrasts (Best, 1995; Flege, 1995; Kuhl and Iverson, 1995; Werker and Curtin, 2005). In none of these theories, however, is the mapping specified in enough detail to allow a concrete implementation. In addition, in most of the existing theories,[1] even if a fully specified mapping was available, it remains unclear how predictions on patterns of error rates could be derived from it (the filtering operation). These theories remain therefore mainly descriptive.

In this paper, we propose to leverage Automatic Speech Recognition (ASR) technology to obtain fully specified mappings between foreign sounds and native categories and then use the *machine ABX* evaluation task (Schatz *et al.*, 2013; Schatz, 2016) to derive quantitative predictions from these mappings regarding cross-linguistic phonetic category perception. More specifically, our approach can be broken down into three steps. First, train a *phoneme recognizer* in a "native" language using annotated continuous speech recordings. Second, use the trained system to derive *perceptual representations* for test stimuli in a foreign language. In this paper, these will be vectors of posterior probabilities over each of the native phonemes. Third, obtain predictions

[a)] Author to whom correspondence should be addressed. Also at: Department of Linguistics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

for perceptual errors by running a *psychophysical test* over these representations for each foreign contrast. *Machine ABX* discrimination tasks will be used for this.

To showcase the possibilities offered by the approach, we look at predictions obtained for three empirically-attested effects in cross-linguistic phonetic category perception. The first two effects are *global* effects that apply to the set of phonetic contrasts in a language as a whole. First, native contrasts tend to be easier to distinguish than non-native ones (Gottfried, 1984). Second, patterns of perceptual confusions are a function of the native language(s): two persons with the same native language tend to confuse the same foreign sounds, which can be different from sounds confused by persons with another native language (Strange, 1995). Thanks to the quantitative and systematic nature of the proposed approach, these effects are straightforward to study. We show that ASR models can account for both of them. Most effects documented in the empirical literature on cross-linguistic phonetic category perception are more *local* however. They describe patterns of confusion observed for very specific choices of languages and contrasts. We illustrate how such effects can be studied with our method through the classical example of AE /ɹ/-/l/ perception by native Japanese listeners (Goto, 1971; Miyawaki *et al.*, 1975). We show that ASR models correctly predict the difficulty of perceiving this distinction for Japanese listeners.

Previous attempts at specifying mappings between foreign and native categories relied on phonological descriptions of the languages involved. Analyses at the level of abstract (context-independent) phonemes, however, were found not to be sufficient to fully account for perceptual data (Kohler, 1981; Strange *et al.*, 2004). For example, the French [u-y] contrast can be either easy or hard to perceive for native AE listeners, depending on the specific phonetic context in which it is realized (Levy and Strange, 2002). Attempting to specify mappings *explicitly* through finer-grain phonetic analyses certainly remains an option, but involves a formidable amount of work. An attractive and potentially less costly alternative consists in specifying mappings *implicitly*, through quantitative models of native speech perception. By this, we mean models that map any input sound to a perceptual representation adapted to the model's "native language." This representation can take the form of a phonetic category label, a vector of posterior probabilities over possible phones or some other, possibly richer, form of representation. Predictions regarding human perception of foreign speech sounds are then derived by analyzing the "native representations" produced by the model when exposed to these foreign sounds.

Let us now explain the rationale for turning toward ASR technology, when the goal is to model *human* speech perception. This approach is best understood in the context of a top-down effort, where the focus is on developing models first at the *information processing* level, before considering issues at the algorithmic and biological implementation levels (Marr, 1982). Native speech perception is thought to arise primarily from a need to reliably identify the linguistic content in the language-specific speech signal to which we are exposed, despite extensive para-linguistic variations. ASR systems, whose goal is to map input speech to corresponding sequences of words, face the same problem. ASR systems seek optimal performance, and can thus be interesting as potential normative models of human behavior from an *efficient coding* point of view (Barlow, 1961), even though biological plausibility is not taken into account in their development.

We found two previous studies taking steps in the proposed direction. In the first one (Strange *et al.*, 2004), a Linear Discriminant Analysis model was trained to classify AE vowels from $F1/F2/F3$ formant plus duration representations. The classification of North German vowels by this model was then compared to assimilation patterns from a phoneme classification task performed by native AE speakers exposed to North German vowels. The model's predictions only partially matched observed human behavior. In the second study (Gong *et al.*, 2010), Hidden-Markov-Models (HMM) with a structure inspired from ASR technology were trained to classify Mandarin consonants from Mel-Frequency Cepstral Coefficients[2] (MFCC). The classification of AE consonants by this model was then compared to assimilation patterns from a phoneme classification task performed by native Mandarin speakers exposed to AE consonants. There was a good consistency between the model's predictions and human assimilation patterns in most cases, although the model provided more variable answers overall and differed markedly from humans in its preferred Mandarin classification of certain AE fricatives.

The present work expands over these previous studies in several respects. First, we replace *ad hoc* speech processing models trained on restricted stimuli[3] with general-purpose ASR systems trained on natural continuous speech. This has both conceptual and practical benefits. Conceptually, the information processing problem our models attempt to solve is closer to the one solved by humans, who have to deal with the full variability of natural speech. From a practical point of view, this allows us to capitalize on existing corpora of annotated speech recordings developed for ASR. A second difference with previous studies is that we improve on the evaluation methodology, by replacing informal analysis of assimilation patterns with quantitative evaluations based on a simple model of an ABX discrimination task, leading to clean and clearly interpretable results. Finally, we conduct more systematic evaluations, testing for two *global* and one *local* effect in cross-linguistic phonetic category perception.

2. Methods

2.1 Speech recordings

To train and evaluate ASR models, five corpora of recorded speech in different languages were used: a subset of the Wall Street Journal (WSJ) corpus (Paul and Baker, 1992), the Buckeye (BUC) corpus (Pitt *et al.*, 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 and Schultz, 2009). Important characteristics of the corpora are summarized in Table 1. Two corpora in AE were included to dissociate *language-mismatch* effects, in which we are interested, from *channel-mismatch* effects due to differences across corpora in recording conditions, microphones, speech register, etc. Phonetic transcriptions were obtained by combining word-level transcriptions with a phonetic dictionary 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 an ASR system similar to those described in Sec. 2.2, but trained on the whole corpus.

2.2 ASR models

State-of-the-art ASR systems are built from deep recurrent neural networks. These systems, however, typically require hundreds of hours of data to be reliably trained and we decided to focus in this study on using older, but more stable, Gaussian-Mixture based Hidden-Markov Models (GMM-HMM) to ensure a reasonable performance across all corpora. Each corpus was randomly split into training and a test set of approximately the same size, each containing an equal number of speakers. There was no overlap between training and test speakers. Models were trained with the Kaldi toolkit (Povey *et al.*, 2011) using the same recipe with the same parameters and input features to train all models.[4] The Word-Error Rate[5] (WER) on the test set for each of the resulting models is reported in Table 1.

We will not attempt to describe the inner workings of the models beyond mentioning that a generative model is trained for each phone, with explicit mechanisms for handling variability due to changes in speaker, phonetic context, or word-position. We refer to the Kaldi documentation for further detail.[6] Input to the models takes the form of 39 MFCC[7] plus 9 pitch-related features[8] extracted every 10 ms of signal. These 48-dimensional input features can be seen as a *universal* auditory-like baseline representation that is not tuned to any particular native language. The model produces native representations under the form of output vectors produced every 10 ms, which list the posterior probabilities, according to the model, that the corresponding stretch of speech signal belongs to each of the segment in the phonemic inventory of the model's native language.[9] The test set of each corpus is decoded with each of the five ASR models and we also use the input features directly, without any GMM-HMM decoding, as a language-independent control, yielding a total of six different representations of each corpus to be evaluated.

2.3 Machine ABX evaluation

We evaluate our ASR models with a machine version of an ABX discrimination task (Schatz *et al.*, 2013; Schatz, 2016) that allows us to quantify how easy it is to

Table 1. WERs 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	143 h	Read	338	8.5%
BUC	AE	19 h	Spont.	40	48.0%
CSJ	Japanese	15 h	Spont.	75	30.0%
GPM	Mandarin	30 h	Read	132	31.0%
GPV	Vietnamese	20 h	Read	129	23.5%

distinguish two phonetic categories based on representations produced by one of our models. The basic idea is to take two acoustic realizations A and X from one of the phonetic categories and one acoustic realization B from the other category and to test whether the model representation for X is closer to the model representation for A than to the model representation for B. The probability for this to be false for A, B, and X randomly chosen in a corpus is defined as the *ABX error rate* for the two phonetic categories according to the model. If it is equal to 0, the two categories are perfectly discriminated. If it is equal to 0.5, discrimination is at chance level.

For each A, B, and X triplet, we use the phone-level time alignments to select corresponding model representations. Because the stimuli have variable durations, the resulting representations can have different lengths. To find a good alignment and obtain a quantitative measure of dissimilarity between A and X and B and X, we use Dynamic Time Warping based on a frame-wise symmetric Kullback-Leibler divergence for posterior probability vectors and a frame-wise cosine distance for the input features control. In the specific ABX task considered here, we select only triplets such that A, B, and X occur in the same phonetic context (same preceding phone and same following phone) and are uttered by the same speaker. For each phonetic contrast an aggregated ABX error rate is obtained by averaging over stimulus order, context, and speaker. Let us illustrate this through the example of the /u/-/i/ contrast. First, we average error rates obtained when A and X are chosen to be /u/ and B is chosen to be /i/ and vice versa, then we average over all possible choices of speaker and finally we average over all possible choices of preceding and following phones. We either report directly the scores obtained for individual phonetic contrasts or we average them over interesting classes of contrasts, such as consonant contrasts or vowel contrasts.

Note that, because we are studying very robust empirical effects that reflect what subjects learn outside the lab and that are expected to be observed in any well-designed experimental task, our evaluation method focuses on simplicity of application rather than detailed modeling of human performance in a specific experimental setting.

3. Results

See the supplementary material,[10] for the raw (unanalyzed) confusion matrices obtained for each model on each test corpus.

3.1 Native vs non-native contrasts

Native phonetic categories are easier to distinguish than non-native categories (Gottfried, 1984). This is consistent with the predictions of our models shown in Fig. 1. The AE models (in red) separate AE phonetic categories better than other models (in blue). This is true even when they are tested with AE stimuli from a corpus different from the one on which they were trained, showing that the differences observed cannot be explained simply by *channel-mismatch* effects and reflect a true *language-specificity* of the representations learned by the models. Another interesting observation is that, while a moderate improvement in phone separability is observed when comparing native AE models to the "universal" input features control, the most salient effect is a large decrease in performance for "non-native" models. A possible interpretation is that, while ASR models can provide categorical representations of native speech that are much more compact than the input features, they do it at the expense of a loss of representation power for coding speech in other languages.[11]

3.2 Native-language-specific confusion patterns

The specific confusions we make between sounds of a foreign language differ according to our native language (Strange, 1995). Consistent with this effect, Fig. 2 shows that, for both consonant and vowel contrasts, the confusion patterns obtained with the two AE models over the different corpora are more similar to each other than to the confusion

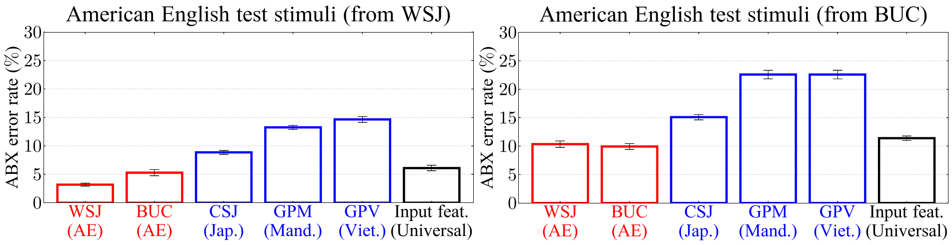

Fig. 1. (Color online) ABX error-rates averaged over all consonant contrasts of AE. Left: using stimuli from the WSJ corpus test set. Right: using stimuli from the BUC corpus test set.

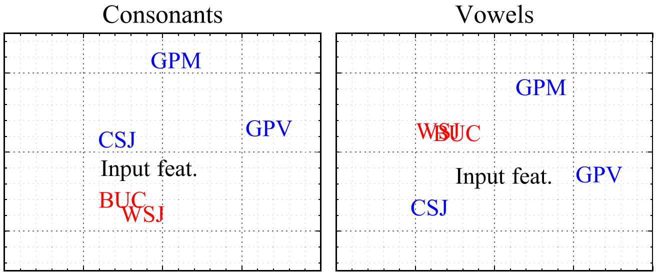

Fig. 2. (Color online) Two-dimensional embeddings of the different models based on the average cosine similarity between their patterns of ABX errors across the five test corpora. The distance between models in the embedding space directly reflects whether they make the same type of confusions or not. Left: for consonant contrasts. Right: for vowel contrasts. Text labels are centered horizontally and vertically on the point they represent.

patterns obtained with models trained on other languages. Confusion patterns for input features occupy a somewhat central role. In this figure, the distance between two points is proportional to the observed similarity between confusion patterns obtained from the associated models.[12] Confusion patterns on a given corpus consist of vectors listing the ABX errors for either all consonant contrasts or all vowel contrasts in this corpus. For example for a language with n consonants, $n(n-1)/2$ consonant contrasts can be formed and the corresponding ABX errors are listed in a vector of size $n(n-1)/2$. The similarity between confusion patterns of two models is defined as the average of the cosine similarity between the confusion patterns obtained with these models on each of the five corpora.[13] Importantly, the rescaling invariance of the cosine similarity ensures that our analysis of confusion patterns is independent from the average ABX error rates studied in Sec. 3.1.

3.3 Japanese listeners and AE /ɹ/-/l/

AE /ɹ/ and /l/ are much harder to perceive for Japanese than for AE native speakers (Goto, 1971; Miyawaki *et al.*, 1975). Figure 3 shows that our models' predictions are fully consistent with this effect: when comparing the Japanese model to both AE models and to the input features, the /ɹ/-/l/ discriminability drops spectacularly, much more than the discriminability of two controls. This is observed when using test stimuli from the WSJ and from the BUC corpora. The first control is the AE /w/-/j/ contrast. Like /ɹ/ and /l/, /w/ and /j/ are liquid consonants, but unlike those, they have a clear counterpart in Japanese. The second control is the average ABX error rate from Sec. 3.1. This control allows to check that there is a specific deficit of the Japanese model on AE /ɹ/-/l/ discrimination, that cannot be explained by an overall weakness of this model.

4. Discussion

Fully specified mappings between foreign sounds and native phonetic categories were obtained for several language pairs through GMM-HMM ASR systems. Coupled with a simple model of a discrimination task, they successfully accounted for several empirically attested effects in cross-linguistic phonetic category perception by monolingual listeners. This includes 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

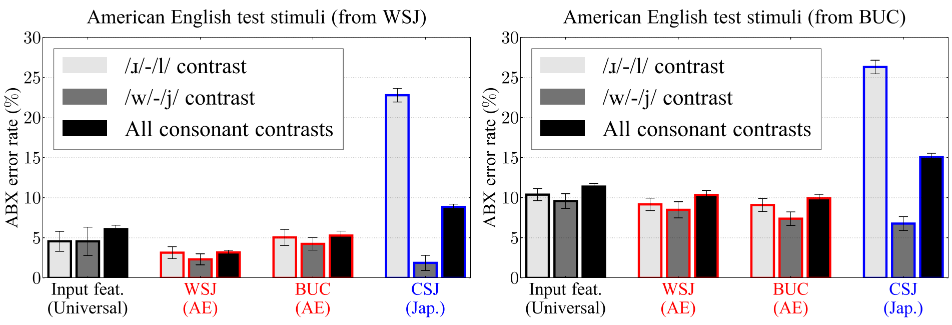

Fig. 3. (Color online) Comparison of the ABX error-rates obtained with the input features, with the two AE models and with the Japanese model on the AE /ɹ/-/l/ contrast. ABX error-rates for the /w/-/j/ contrast and ABX error-rates averaged over all consonant contrasts of AE are also shown as controls. Left: using stimuli from the WSJ corpus test set. Right: using stimuli from the BUC corpus test set.

do not make the same confusions between phonetic categories of AE than native speakers of French). We also showed that the proposed model can account for a well-known *local* effect: AE /ɹ/ and /l/ are very hard to discriminate for native speakers of Japanese.

These results provide a proof-of-concept for the proposed approach to evaluating ASR systems as quantitative models of phonetic category perception. They also show promise regarding the possibility of modeling human phonetic category perception with ASR systems. Yet we do not claim, at this point, to have provided definitive evidence that the particular GMM-HMM ASR systems considered provide the best, or even a particularly "good," such model. A host of *local* effects have been documented in the empirical literature on phonetic category perception beyond the one investigated here (Strange, 1995; Cutler, 2012) and the empirical adequacy of the proposed models with respect to more of these effects will need to be determined before any conclusion can be reached. Effects that are hard to predict from conventional phonological analyses, such as how the phonetic or prosodic context can modulate the difficulty of perceiving certain foreign contrasts (Levy and Strange, 2002; Kohler, 1981; Strange *et al.*, 2004), should be of particular interest. Finally, let us underline that we only investigated predictions obtained with one particular ASR architecture. There are multiple ways of instantiating ASR systems, which might yield different predictions. For example, modeling variability in the signal due to the phonetic context explicitly with context-dependent phone models, as in this article, or implicitly with context-independent phone models, might affect predictions regarding the aforementioned context-dependent effects. Another example of a potentially significant decision is whether to use HMM-GMM or neural-network systems. HMM models have known structural limitations for modeling segment duration (Pylkkönen and Kurimo, 2004), from which neural-network models do not suffer. Thus, neural-network ASR systems may provide better models of native perception in languages like Japanese, where duration is contrastive. The multiplicity of documented empirical effects and available computational models calls for an extensive investigation, which could in turn trigger a more systematic *experimental* investigation of non-native perception and result in applications in foreign language education.

Acknowledgments

The authors would like to thank Naomi Feldman, Alex Cristia, and anonymous reviewers for helpful comments on earlier versions of this manuscript. The research leading to these results received funding from the European Research Council under the FP/2007-2013 program/ERC Grant Agreement No. ERC-2011-AdG-295810 BOOTPHON, from the Agence Nationale pour la Recherche (Grant Nos. ANR-2010-BLAN-1901-1 BOOTLANG, ANR-10-0001-02 PSL*, and ANR-10-LABX-0087 IEC), from the Fondation de France and from the National Science Foundation (Grant No. BCS-1734245). Research leading to this publication was performed while T.S. was a Ph.D. student in the Laboratoire de Sciences Cognitives et Psycholinguistique at École Normale Supérieure in Paris. This paper was written in part while E.D. was a visiting researcher at Facebook AI Research, Paris and while T.S. was a visiting researcher in the Computational Psycholinguistics Laboratory at the Massachusetts Institute of Technology.

References and links

[1] Best (1995) being a possible exception.

[2] MFCC (Mermelstein, 1976) are speech features commonly used as a front-end to ASR systems. They can be thought of as moderate-dimensional descriptor ($d = 13$) of the whole shape of regularly-spaced spectral-slices in a Mel-scale log-spectrogram. They are usually taken every 10 ms and augmented with their first and second time derivatives to incorporate dynamic information, leading to 100 vector descriptors of dimension $d = 39$ per second of signal.

[3] Previous studies used as training stimuli a limited sample of 264 AE vowels occurring either in [hVba] context or within a unique carrier sentence (Strange *et al.*, 2004) and 3331 Chinese consonants occurring in isolated VCV context (Gong *et al.*, 2010).

[4] See https://goo.gl/RsKMA3.

[5] Error-rate obtained in a word recognition task using the trained acoustic model with a language model (in our case a word-level bigram estimated from the training set).

[6] See http://kaldi-asr.org/.

[7] See footnote 1.

[8] Pitch features were added because two of the languages considered (Mandarin and Vietnamese) are tonal languages.

[9] More specifically, we use Viterbi-smoothed phone-level posterior grams obtained with a phone-level bigram language model estimated on the training set of each corpus.

[10] See supplementary material at https://doi.org/10.1121/1.5037615 for the raw (unanalyzed) confusion matrices obtained for each model on each test corpus.

[11] Note that Renshaw *et al.* (2015) observed a different pattern when testing a neural-network-based ASR system trained on AE on the Xitsonga language: the "AE-native" model improved Xitsonga phone separability relative to the input features control. There are, at least, two possible interpretations for this discrepancy: it could be due to general differences between GMM-HMM and neural-network architectures or it could be due to differences in the representation format chosen (they used "bottleneck features" extracted from a middle layer of the neural network, which are not constrained to represent phonetic categories, while our posterior features are).

[12] Two-dimensional embeddings are obtained with scikit-learn's non-metric multi-dimensional-scaling.

[13] Observed range of cosine similarities: [0.90–0.96] for consonants and [0.85–0.94] for vowels.

Barlow, H. B. (**1961**). *Possible Principles Underlying the Transformations of Sensory Messages* (MIT Press, Cambridge, MA).

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

Cutler, A. (**2012**). *Native Listening: Language Experience and the Recognition of Spoken Words* (MIT Press, Cambridge, MA).

Flege, J. E. (**1995**). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (York Press, York, England), pp. 233–277.

Gong, J., Cooke, M., and Garcia Lecumberri, M. (**2010**). "Towards a quantitative model of mandarin Chinese perception of English consonants," in *Proceedings of New Sounds*.

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," J. Phonetics **12**(2), 91–114.

Kohler, K. (**1981**). "Contrastive phonology and the acquisition of phonetic skills," Phonetica **38**(4), 213–226.

Kuhl, P. K., and Iverson, P. (**1995**). "Linguistic experience and the 'perceptual magnet effect,'" in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (York Press, York, England), pp. 121–154.

Levy, E. S., and Strange, W. (**2002**). "Effects of consonantal context on perception of French rounded vowels by American English adults with and without French language experience," J. Acoust. Soc. Am. **111**(5), 2361–2362.

Maekawa, K. (**2003**). "Corpus of spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Marr, D. (**1982**). *Vision: A Computational Approach* (Freeman, San Francisco, CA).

Mermelstein, P. (**1976**). "Distance measures for speech recognition, psychological and instrumental," Pattern Recogn. Artif. Intell. **116**, 91–103.

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

Paul, D. B., and Baker, J. M. (**1992**). "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, pp. 357–362.

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

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., and Veselý, K. (**2011**). "The Kaldi speech recognition toolkit," in *Proceedings of the Workshop on Automatic Speech Recognition and Understanding*.

Pylkkönen, J., and Kurimo, M. (**2004**). "Duration modeling techniques for continuous speech recognition," in *Proceedings of INTERSPEECH*.

Renshaw, D., Kamper, H., Jansen, A., and Goldwater, S. (**2015**). "A comparison of neural network methods for unsupervised representation learning on the Zero Resource Speech Challenge," in *Proceedings of 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., and Dupoux, E. (**2013**). "Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline," in *Proceedings of INTERSPEECH*.

Schultz, T. (**2002**). "Globalphone: A multilingual speech and text database developed at Karlsruhe University," in *Proceedings of INTERSPEECH*.

Strange, W. (**1995**). *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (York Press, York, England).

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

Vu, N. T., and Schultz, T. (**2009**). "Vietnamese large vocabulary continuous speech recognition," in *Proceedings of ASRU*.

Werker, J. F., and Curtin, S. (**2005**). "Primir: A developmental framework of infant speech processing," Lang. Learn. Develop. **1**(2), 197–234.