

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

Thomas Schatz, Francis Bach, Emmanuel Dupoux

► To cite this version:

Thomas Schatz, Francis Bach, Emmanuel Dupoux. Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception. *Journal of the Acoustical Society of America*, Acoustical Society of America, 2018, 143 (5), pp.EL372 - EL378. 10.1121/1.5037615 . hal-01888735

HAL Id: hal-01888735

<https://hal.archives-ouvertes.fr/hal-01888735>

Submitted on 7 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Thomas Schatz

*Department of Linguistics & UMIACS
University of Maryland
College Park, USA*

*Department of Linguistics
Massachusetts Institute of Technology
Cambridge, USA*

*LSCP, Département d'études cognitives de l'ENS
École normale supérieure, EHESS, CNRS, PSL Research University
29, rue d'Ulm 75005 Paris, France
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*

Quantitative models of phonetic category perception

1
2
3
4
5
6
7
8
9
10
11
12
13

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 (ASR) systems trained on continuous speech corpora are used to provide a fully specified mapping between foreign sounds and native categories. We show how the *machine ABX* evaluation method can be used to compare predictions from the resulting quantitative models with empirically at-tested effects in human cross-linguistic phonetic category perception.

© 2018 Acoustical Society of America

Keywords: phonetic categories; human perception; quantitative modeling; ASR; machine ABX.

14 1. Introduction

15 The way we perceive phonetic categories (i.e. basic speech sounds such as consonants
16 and vowels) is largely determined by the language(s) to which we were exposed as
17 a child. For example, native speakers of Japanese have a hard time discriminating
18 between American English (AE) /ɪ/ and /I/, a phonetic contrast that has no equiva-
19 lent in Japanese (Goto, 1971; Miyawaki et al., 1975). Perceptual specialization to the
20 phonological properties of the native language has been extensively investigated using
21 a varieties of techniques (see Strange 1995 and Cutler 2012 for reviews). Many of the
22 proposed theoretical accounts of this phenomenon concur that foreign sounds are not
23 perceived faithfully, but rather, are ‘mapped’ onto one’s pre-existing (native) phonetic
24 categories, which act as a kind of ‘filter’ resulting in the degradation of some non-
25 native contrasts (Best, 1995; Flege, 1995; Kuhl and Iverson, 1995; Werker and Curtin,
26 2005). In none of these theories, however, is the mapping specified in enough detail to
27 allow a concrete implementation. In addition, in most of the existing theories¹, even if
28 a fully specified mapping was available, it remains unclear how predictions on patterns
29 of error rates could be derived from it (the filtering operation). These theories remain
30 therefore mainly descriptive.

31 In this paper, we propose to leverage ASR technology to obtain fully speci-
32 fied mappings between foreign sounds and native categories and then use the *machine*
33 *ABX* evaluation task (Schatz et al., 2013; Schatz, 2016) to derive quantitative pre-
34 dictions from these mappings regarding cross-linguistic phonetic category perception.
35 More specifically, our approach can be broken down into three steps. First, train a
36 *phoneme recognizer* in a ‘native’ language using annotated continuous speech record-
37 ings. Second, use the trained system to derive *perceptual representations* for test stimuli
38 in a foreign language. In this paper, these will be vectors of posterior probabilities over
39 each of the native phonemes. Third, obtain predictions for perceptual errors by run-
40 ning a *psychophysical test* over these representations for each foreign contrast. *Machine*
41 *ABX* discrimination tasks will be used for this.

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

58 Previous attempts at specifying mappings between foreign and native cate-
59 gories relied on phonological descriptions of the languages involved. Analyses at the
60 level of abstract (context-independent) phonemes, however, were found not to be suf-
61 ficient to fully account for perceptual data (Kohler, 1981; Strange et al., 2004). For
62 example, the French [u-y] contrast can be either easy or hard to perceive for native AE
63 listeners, depending on the specific phonetic context in which it is realized (Levy and
64 Strange, 2002). Attempting to specify mappings *explicitly* through finer-grain phonetic
65 analyses certainly remains an option, but involves a formidable amount of work. An

66 attractive and potentially less costly alternative consists in specifying mappings *implicit-*
67 *ly*, through quantitative models of native speech perception. By this, we mean models
68 that map any input sound to a perceptual representation adapted to the model's 'native
69 language'. This representation can take the form of a phonetic category label, a vector
70 of posterior probabilities over possible phones or some other, possibly richer, form of
71 representation. Predictions regarding human perception of foreign speech sounds are
72 then derived by analyzing the 'native representations' produced by the model when
73 exposed to these foreign sounds.

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

86 We found two previous studies taking steps in the proposed direction. In the
87 first one (Strange et al., 2004), a Linear Discriminant Analysis model was trained to
88 classify AE vowels from F1/F2/F3 formant plus duration representations. The classi-
89 fication of North German vowels by this model was then compared to assimilation
90 patterns from a phoneme classification task performed by native AE speakers exposed
91 to North German vowels. The model's predictions only partially matched observed hu-
92 man behavior. In the second study (Gong et al., 2010), Hidden-Markov-Models (HMM)
93 with a structure inspired from ASR technology were trained to classify Mandarin con-
94 sonants from Mel-Frequency Cepstral Coefficients² (MFCC). The classification of AE
95 consonants by this model was then compared to assimilation patterns from a phoneme
96 classification task performed by native Mandarin speakers exposed to AE consonants.
97 There was a good consistency between model's predictions and human assimilation
98 patterns in most cases, although the model provided more variable answers overall
99 and differed markedly from humans in its preferred Mandarin classification of certain
100 AE fricatives.

101 The present work expands over these previous studies in several respects. First,
102 we replace ad hoc speech processing models trained on restricted stimuli³ with general-
103 purpose ASR systems trained on natural continuous speech. This has both conceptual
104 and practical benefits. Conceptually, the information processing problem our models
105 attempt to solve is closer to the one solved by humans, who have to deal with the full
106 variability of natural speech. From a practical point of view, this allows us to capital-
107 ize on existing corpora of annotated speech recordings developed for ASR. A second
108 difference with previous studies is that we improve on the evaluation methodology,
109 by replacing informal analysis of assimilation patterns with quantitative evaluations
110 based on a simple model of an ABX discrimination task, leading to clean and clearly
111 interpretable results. Finally, we conduct more systematic evaluations, testing for two
112 *global* and one *local* effect in cross-linguistic phonetic category perception.

113 2. Methods

114 2.1. Speech recordings

115 To train and evaluate ASR models, 5 corpora of recorded speech in different languages
116 were used: a subset of the Wall Street Journal corpus (WSJ) (Paul and Baker, 1992),

117 the Buckeye corpus (BUC) (Pitt et al., 2005), a subset of the Corpus of Spontaneous
118 Japanese (CSJ) (Maekawa, 2003), the Global Phone Mandarin (GPM) corpus (Schultz,
119 2002) and the Global Phone Vietnamese (GPV) corpus (Vu and Schultz, 2009). Important
120 characteristics of the corpora are summarized in Table 1. Two corpora in American
121 English were included to dissociate *language-mismatch* effects, in which we are inter-
122 ested, from *channel-mismatch* effects due to differences across corpora in recording
123 conditions, microphones, speech register, etc. Phonetic transcriptions were obtained
124 by combining word-level transcriptions with a phonetic dictionary for the WSJ, BUC,
125 GPM and GPV corpora. For the CSJ corpus, manual phonetic transcriptions were used.
126 For all corpora, timestamps for the phonetic transcriptions were obtained by forced
127 alignment using an ASR system similar to those described in the next section, but
128 trained on the whole corpus.

129 2.2. ASR models

130 State-of-the-art ASR systems are built from deep recurrent neural networks. These sys-
131 tems, however, typically require hundreds of hours of data to be reliably trained and
132 we decided to focus in this study on using older, but more stable, Gaussian-Mixture
133 based Hidden-Markov Models (GMM-HMM) to ensure reasonable performance across
134 all corpora. Each corpus was randomly split into a training and a test set of approx-
135 imately the same size, each containing an equal number of speakers. There was no
136 overlap between training and test speakers. Models were trained with the Kaldi toolkit
137 (Povey et al., 2011) using the same recipe with the same parameters and input fea-
138 tures to train all models⁴. The Word-Error Rate⁵ (WER) on the test set for each of the
139 resulting models is reported in Table 1.

140 We will not attempt to describe the inner workings of the models beyond men-
141 tioning that a generative model is trained for each phone, with explicit mechanisms for
142 handling variability due to changes in speaker, phonetic context or word-position. We
143 refer to the Kaldi documentation for further detail⁶. Input to the models takes the form
144 of 39 MFCC coefficients⁷ plus 9 pitch-related features⁸ extracted every 10ms of signal.
145 These 48-dimensional input features can be seen as a *universal* auditory-like baseline
146 representation that is not tuned to any particular ‘native language’. The model pro-
147 duces ‘native’ representations under the form of output vectors produced every 10ms,
148 which list the posterior probabilities, according to the model, that the corresponding
149 stretch of speech signal belongs to each of the segment in the phonemic inventory of
150 the model’s ‘native language’⁹. The test set of each corpus is decoded with each of the
151 5 ASR models and we also use the input features directly, without any GMM-HMM
152 decoding, as a language-independent control, yielding a total of 6 different represen-
153 tations of each corpus to be evaluated.

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%

154 *2.3. Machine ABX evaluation*

155 We evaluate our ASR models with a machine version of an ABX discrimination task
156 (Schatz et al., 2013; Schatz, 2016) that allows us to quantify how easy it is to distin-
157 guish two phonetic categories based on representations produced by one of our models.
158 The basic idea is to take two acoustic realizations A and X from one of the phonetic
159 categories and one acoustic realization B from the other category and to test whether
160 the model representation for X is closer to the model representation for A than to
161 the model representation for B . The probability for this to be false for A , B and X
162 randomly chosen in a corpus is defined as the *ABX error rate* for the two phonetic
163 categories according to the model. If it is equal to 0, the two categories are perfectly
164 discriminated. If it is equal to .5, discrimination is at chance level.

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

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

185 **3. Results**

186 See supplementary material for the raw (unanalyzed) confusion matrices obtained for
187 each model on each test corpus.

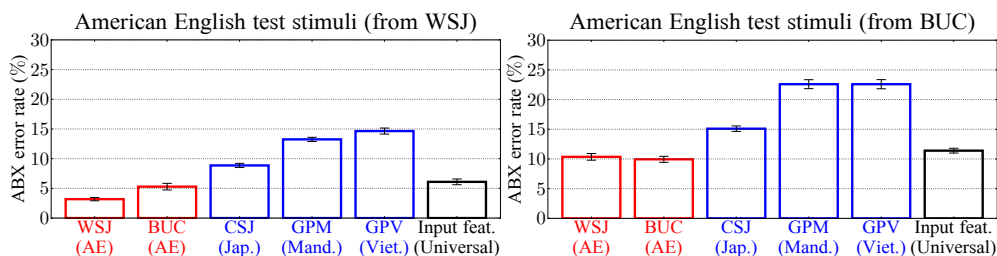
188 *3.1. Native vs. non-native contrasts*

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.

189 Native phonetic categories are easier to distinguish than non-native categories
190 (Gottfried, 1984). This is consistent with the predictions of our models shown in Figure

191 1. The AE models (in red) separate AE phonetic categories better than other models (in
 192 blue). This is true even when they are tested with AE stimuli from a corpus different
 193 from the one on which they were trained, showing that the differences observed cannot
 194 be explained simply by *channel-mismatch* effects and reflect a true *language-specificity*
 195 of the representations learned by the models. Another interesting observation is that,
 196 while a moderate improvement in phone separability is observed when comparing
 197 ‘native’ AE models to the ‘universal’ input features control, the most salient effect is
 198 a large decrease in performance for ‘non-native’ models. A possible interpretation is
 199 that, while ASR models can provide categorical representations of ‘native’ speech that
 200 are much more compact than the input features, they do it at the expense of a loss of
 201 representation power for coding speech in other languages¹⁰.

202 3.2. Native-language-specific confusion patterns

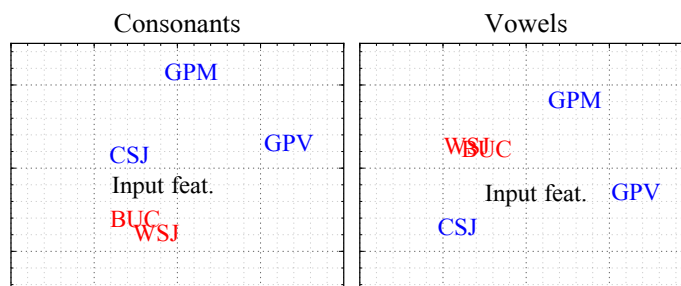


Fig. 2. (color online) Two-dimensional embeddings of the different models based on the average cosine similarity between their patterns of ABX errors across the five test corpora. The distance between models in the embedding space directly reflects whether they make the same type of confusions or not.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

203 The specific confusions we make between sounds of a foreign language differ
 204 according to our native language (Strange, 1995). Consistent with this effect, Figure 2
 205 shows that, for both consonant and vowel contrasts, the confusion patterns obtained
 206 with the two AE models over the different corpora are more similar to each other than
 207 to the confusion patterns obtained with models trained on other languages. Confusion
 208 patterns for input features occupy a somewhat central role. In this figure, the distance
 209 between two points is proportional to the observed similarity between confusion pat-
 210 terns obtained from the associated models¹¹. Confusion patterns on a given corpus
 211 consist of vectors listing the ABX errors for either all consonant contrasts or all vowel
 212 contrasts in this corpus. For example for a language with n consonants, $n(n - 1)/2$
 213 consonant contrasts can be formed and the corresponding ABX errors are listed in a
 214 vector of size $n(n - 1)/2$. The similarity between confusion patterns of two models is
 215 defined as the average of the cosine similarity between the confusion patterns obtained
 216 with these models on each of the five corpora¹². Importantly, the rescaling invariance
 217 of the cosine similarity ensures that our analysis of confusion patterns is independent
 218 from the average ABX error rates studied in Section 3.1.

219 3.3. Japanese listeners and American English /ɪ/-/I/

220 AE /ɪ/ and /I/ are much harder to perceive for Japanese than for AE native speak-
 221 ers (Goto, 1971; Miyawaki et al., 1975). Figure 3 shows that our models’ predictions
 222 are fully consistent with this effect: when comparing the Japanese model to both AE
 223 models and to the input features, the /ɪ/-/I/ discriminability drops spectacularly, much

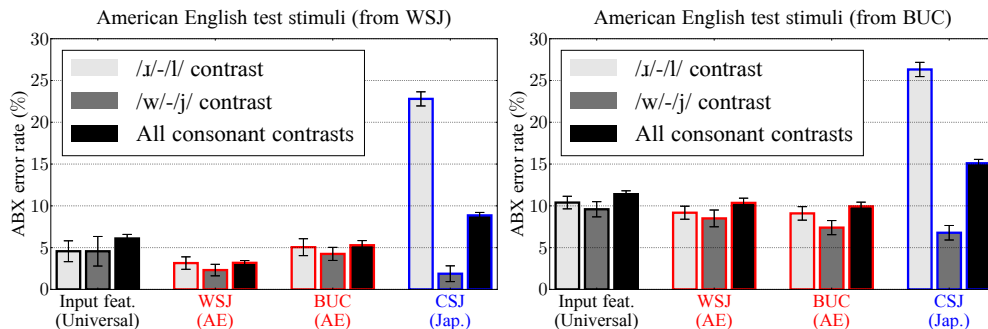


Fig. 3. (color online) Comparison of the ABX error-rates obtained with the input features, with the two AE models and with the Japanese model on the AE /ɹ/-/l/ contrast. ABX Error-rates for the /w/-/j/ contrast and ABX Error-rates averaged over all consonant contrasts of AE are also shown as controls.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

224 more than the discriminability of two controls. This is observed both when using test
225 stimuli from the WSJ and from the BUC corpora. The first control is the AE /w/-/j/
226 contrast. Like /ɹ/ and /l/, /w/ and /j/ are liquid consonants, but unlike those, they have
227 a clear counterpart in Japanese. The second control is the average ABX error rate from
228 Section 3.1. This control allows to check that there is a specific deficit of the Japanese
229 model on AE /ɹ/-/l/ discrimination, that cannot be explained by an overall weakness
230 of this model.

231 4. Discussion

232 Fully specified mappings between foreign sounds and native phonetic categories were
233 obtained for several language pairs through GMM-HMM ASR systems. Coupled with a
234 simple model of a discrimination task, they successfully accounted for several empirically
235 attested effects in cross-linguistic phonetic category perception by monolingual
236 listeners. This includes two types of *global* effects: first, that the phonetic categories
237 of a language are overall harder to discriminate for non-native speakers than for native
238 speakers and second, that the pattern of confusions between phonetic categories
239 for non-native speakers is specific to their native language (e.g. native speakers of
240 Japanese do not make the same confusions between phonetic categories of American
241 English than native speakers of French). We also showed that the proposed model can
242 account for a well-known *local* effect: American English /ɹ/ and /l/ are very hard to
243 discriminate for native speakers of Japanese.

244 These results provide a proof-of-concept for the proposed approach to evaluat-
245 ing ASR systems as quantitative models of phonetic category perception. They also
246 show promise regarding the possibility of modeling human phonetic category percep-
247 tion with ASR systems. Yet we do not claim, at this point, to have provided definitive
248 evidence that the particular GMM-HMM ASR systems considered provide the best, or
249 even a particularly 'good', such model. A host of *local* effects have been documented
250 in the empirical literature on phonetic category perception beyond the one investi-
251 gated here (Strange, 1995; Cutler, 2012) and the empirical adequacy of the proposed
252 models with respect to more of these effects will need to be determined before any
253 conclusion can be reached. Effects that are hard to predict from conventional phono-
254 logical analyses, such as how the phonetic or prosodic context can modulate the dif-
255 ficulty of perceiving certain foreign contrasts (Levy and Strange, 2002; Kohler, 1981;
256 Strange et al., 2004), should be of particular interest. Finally, let us underline that
257 we only investigated predictions obtained with one particular ASR architecture. There

258 are multiple ways of instantiating ASR systems, which might yield different predic-
259 tions. For example, modeling variability in the signal due to the phonetic context
260 explicitly with context-dependent phone models, as in this article, or implicitly with
261 context-independent phone models, might affect predictions regarding the aforemen-
262 tioned context-dependent effects. Another example of a potentially significant decision
263 is whether to use HMM-GMM or neural-network systems. HMM models have known
264 structural limitations for modeling segment duration (Pylkkönen and Kurimo, 2004),
265 from which neural-network models do not suffer. Thus, neural-network ASR systems
266 may provide better models of native perception in languages like Japanese, where du-
267 ration is contrastive. The multiplicity of documented empirical effects and available
268 computational models calls for an extensive investigation, which could in turn trigger
269 a more systematic *experimental* investigation of non-native perception and result in
270 applications in foreign language education.

271 Acknowledgments

272 The authors would like to thank Naomi Feldman, Alex Cristia and anonymous review-
273 ers for helpful comments on earlier versions of this manuscript. The research leading to
274 these results received funding from the European Research Council under the FP/2007-
275 2013 program / ERC Grant Agreement n. ERC-2011-AdG-295810 BOOTPHON, from
276 the Agence Nationale pour la Recherche (ANR-2010-BLAN-1901-1 BOOTLANG, ANR-
277 10-0001-02 PSL*, ANR-10-LABX-0087 IEC), from the Fondation de France and from
278 the National Science Foundation (BCS-1734245). This paper was written in part while
279 Emmanuel Dupoux was a visiting researcher at Facebook AI Research, Paris and while
280 Thomas Schatz was a visiting researcher in the Computational Psycholinguistics Labo-
281 ratory at the Massachusetts Institute of Technology.

282 Notes

283 ¹Best 1995 being a possible exception.

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

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

292 ⁴See <https://goo.gl/RsKMA3>.

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

295 ⁶See <http://kaldi-asr.org/>.

296 ⁷See footnote 1.

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

299 ⁹More specifically, we use Viterbi-smoothed phone-level posteriorgrams obtained with a phone-level
300 bigram language model estimated on the training set of each corpus.

301 ¹⁰Note that Renshaw et al. (2015) observed a different pattern when testing a neural-network-based
302 ASR system trained on AE on the Xitsonga language: the 'AE-native' model improved Xitsonga phone sep-
303 arability relative to the input features control. There are, at least, two possible interpretations for this dis-
304 crepancy: it could be due to general differences between GMM-HMM and neural-network architectures or
305 it could be due to differences in the representation format chosen (they used 'bottleneck features' extracted
306 from a middle layer of the neural network, which are not constrained to represent phonetic categories, while
307 our posterior features are)

308 ¹¹Two-dimensional embeddings are obtained with scikit-learn's non-metric multi-dimensional-scaling.

309 ¹²Observed range of cosine similarities: [0.90-0.96] for consonants and [0.85-0.94] for vowels.

310 References and links

- 311 H. B. Barlow. *Possible principles underlying the transformations of sensory messages*. MIT press, 1961.
312 C. T. Best. A direct realist view of cross-language speech perception. *Speech Perception and Linguistic*
313 *Experience: Issues in Cross-Language Research*, pages 171–204, 1995.

- 314 A. Cutler. *Native listening: Language experience and the recognition of spoken words*. Mit Press, 2012.
- 315 J. E. Flége. Second language speech learning: Theory, findings, and problems. *Speech perception and*
316 *linguistic experience: Issues in cross-language research*, pages 233–277, 1995.
- 317 J. Gong, M. Cooke, and M. Garcia Lecumberri. Towards a quantitative model of mandarin chinese
318 perception of english consonants. *Proc. NewSounds 2010*, 2010.
- 319 H. Goto. Auditory perception by normal japanese adults of the sounds l and r. *Neuropsychologia*, 9
320 (3):317–323, 1971.
- 321 T. L. Gottfried. Effects of consonant context on the perception of french vowels. *Journal of Phonetics*,
322 12(2):91–114, 1984.
- 323 K. Kohler. Contrastive phonology and the acquisition of phonetic skills. *Phonetica*, 38(4):213–226,
324 1981.
- 325 P. K. Kuhl and P. Iverson. Linguistic experience and the perceptual magnet effect. *Speech perception*
326 *and linguistic experience: Issues in cross-language research*, pages 121–154, 1995.
- 327 E. S. Levy and W. Strange. Effects of consonantal context on perception of french rounded vowels by
328 american english adults with and without french language experience. *The Journal of the Acoustical*
329 *Society of America*, 111(5):2361–2362, 2002.
- 330 K. Maekawa. Corpus of spontaneous japanese: Its design and evaluation. In *Proc. ISCA & IEEE*
331 *Workshop on Spontaneous Speech Processing and Recognition*, 2003.
- 332 D. Marr. *Vision: A computational approach*. Freeman.[aAC], 1982.
- 333 P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern*
334 *recognition and artificial intelligence*, 116:91–103, 1976.
- 335 K. Miyawaki, J. J. Jenkins, W. Strange, A. M. Liberman, R. Verbrugge, and O. Fujimura. An effect of
336 linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english.
337 *Perception & Psychophysics*, 18(5):331–340, 1975.
- 338 D. B. Paul and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proc. Workshop*
339 *on Speech and Natural Language*, pages 357–362, 1992.
- 340 M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond. The buckeye corpus of conversational
341 speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–
342 95, 2005.
- 343 D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlčėk,
344 Y. Qian, P. Schwarz, et al. The kaldi speech recognition toolkit. In *Proc. Workshop on Automatic*
345 *Speech Recognition and Understanding*, 2011.
- 346 J. Pyłkkönen and M. Kurimo. Duration modeling techniques for continuous speech recognition. In
347 *Proc. INTERSPEECH*, 2004.
- 348 D. Renshaw, H. Kamper, A. Jansen, and S. Goldwater. A comparison of neural network methods for
349 unsupervised representation learning on the zero resource speech challenge. In *Proc. INTERSPEECH*,
350 2015.
- 351 T. Schatz. *ABX-Discriminability Measures and Applications*. Doctoral dissertation, Université Paris 6
352 (UPMC), 2016.
- 353 T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux. Evaluating speech features
354 with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *Proc. INTERSPEECH*,
355 2013.
- 356 T. Schultz. Globalphone: a multilingual speech and text database developed at karlsruhe university.
357 In *Proc. INTERSPEECH*, 2002.
- 358 W. Strange. *Speech perception and linguistic experience: Issues in cross-language research*. York Press,
359 1995.
- 360 W. Strange, O.-S. Bohn, S. A. Trent, and K. Nishi. Acoustic and perceptual similarity of north german
361 and american english vowels. *The Journal of the Acoustical Society of America*, 115(4):1791–1807,
362 2004.
- 363 N. T. Vu and T. Schultz. Vietnamese large vocabulary continuous speech recognition. In *Proc. ASRU*,
364 2009.
- 365 J. F. Werker and S. Curtin. Primir: A developmental framework of infant speech processing. *Language*
366 *learning and development*, 1(2):197–234, 2005.