A quantitative model of the language familiarity effect in infancy

Craig A. Thorburn (craigtho@umd.edu)
Naomi H. Feldman (nhf@umd.edu)
Thomas Schatz (thomas.schatz.1986@gmail.com)
Department of Linguistics & UMIACS, University of Maryland, College Park, USA

Abstract

Human listeners are better at telling apart speakers of their native language than speakers of other languages, a phenomenon known as the *language familiarity effect*. The recent observation of such an effect in infants as young as 4.5 months of age (Fecher & Johnson, in press) has led to new difficulties for theories of the effect. On the one hand, retaining classical accounts—which rely on sophisticated knowledge of the native language (Goggin, Thompson, Strube, & Simental, 1991)–requires an explanation of how infants could acquire this knowledge so early. On the other hand, letting go of these accounts requires an explanation of how the effect could arise in the absence of such knowledge. In this paper, we build on algorithms from unsupervised machine learning and zero-resource speech technology to propose, for the first time, a feasible acquisition mechanism for the language familiarity effect in infants. Our results show how, without relying on sophisticated linguistic knowledge, infants could develop a language familiarity effect through statistical modeling at multiple time-scales of the acoustics of the speech signal to which they are exposed.

Keywords: language familiarity; language acquisition; unsupervised learning; speech; i-vector

Introduction

Listeners are highly skilled at identifying who is talking on the sole basis of their voice, and are better at doing so in their native language than in an unfamiliar language, a phenomenon known as the *language familiarity effect* (Goggin et al., 1991). This effect was initially thought to require language understanding, a view that has been supported by a number of empirical results. In particular, whereas moderate proficiency in a language is sufficient to support the effect (Köster & Schiller, 1997), simple passive exposure to speech recordings in that language (Perrachione & Wong, 2007) or native proficiency with a typologically related language (Köster & Schiller, 1997) do not seem to help. The effect has also been shown not to be driven by language-specific prosodic cues (Schiller, Koster, & Duckworth, 1997). More recently, it has been proposed that abstract phonological knowledge of the native language might suffice to explain these observations, without requiring language understanding (Johnson, Bruggeman, & Cutler, 2018). Under both accounts, however, the effect relies on sophisticated linguistic knowledge of the target language.

It has thus been surprising to find a similar language familiarity effect in 7.5-month-old infants (Johnson, Westrek, Nazzi, & Cutler, 2011; Fecher & Johnson, 2018) and even 4.5-month-old infants (Fecher & Johnson, in press), whose language comprehension abilities and phonological knowledge are thought to be very limited (Bergelson & Swingley, 2012; Gervain & Mehler, 2010). In these speaker discrimination experiments, infants were habituated to a small number of voices ($n = 1$ or 3) producing utterances either in their native language (native condition) or in a foreign language (non-native condition) and then tested on their ability to detect a new voice producing an utterance in the same language. They only showed evidence of detecting the change in the native condition.

To account for this new discovery, one possibility is to assume that the effect in infants relies on the same mechanism as the one in adults. However, this would imply that infants have more advanced knowledge of their native language than previously believed at this age and would require explaining how they are able to acquire this knowledge so early. Alternatively, Fecher, Paquette-Smith, and Johnson (2019) proposed that the effect found in infants might be a precursor of the effect found in adults, which could emerge in the absence of sophisticated linguistic knowledge. However, the question of what learning mechanism could plausibly lead an infant to develop a language familiarity effect on the basis of very limited linguistic knowledge has not received a satisfactory answer so far.

In this paper, we build on techniques from unsupervised machine learning and speech technology to present a plausible computational model of the language familiarity effect in infants. We show that the effect can arise from modeling speech acoustics in a given language (the model's 'native' language) at multiple time scales without supervision. This demonstrates that it is feasible for infants to develop a language familiarity effect without relying on sophisticated linguistic knowledge, giving credence to the view developed in Fecher et al. (2019) that the language familiarity effect documented in infants might be a less sophisticated precursor of the one studied in adults.

Approach

Our objective is to propose a learning mechanism that could plausibly be at play in young infants, and test whether this mechanism can account for infants' behavior in experiments that have documented the language familiarity effect. We operationalize the question as follows: a learning algorithm—representing the learning mechanism at play in infants—is first trained on raw, untranscribed unsegmented speech recordings in a 'native language'—representing language input plausibly available to an infant. Through this procedure, we train two 'American English native' and two 'Japanese native' models. Representations of test utterances from various speakers in either language are then extracted from each model and

Table 1: Language, training and test set duration, speech register, and number of speakers in the training and test sets for each corpus.

Corpus	Language	Train	Test	Register	Train Speakers	Test Speakers
WSJ	American English	19h30	9h39	Read	95	47
GPJ	Japanese	19h33	9h40	Read	95	47
BUC	American English	9h13	9h01	Spontaneous	20	20
CSJ	Japanese	9h11	8h57	Spontaneous	20	19

used to compute a measure of speaker discriminability in each language—which stands in for the laboratory assessment of infant's speaker discrimination abilities. If we find that speakers are easier to discriminate in each model's 'native' language, this would mean that the learning algorithm is successful in accounting for the language familiarity effect in infants.

To assess the robustness of the results and control for confounds, we perform model training and speaker discrimination tests on corpora of two registers—either spontaneous or read speech in each language. Testing on both a native and non-native language of a different register allows us to check that an effect is the result of the language the model is trained on, rather than idiosyncratic properties of a particular corpus.

The learning mechanism we propose combines modeling at the level of speech frames (25 ms-long windows of speech signal sampled every 10 ms) with an utterance-level adaptation mechanism, which learns a model of the utterance-to-utterance variability in the parameters of the frame-level model. This is motivated by theories of early language acquisition and insights from speech technology. In particular, the proposed algorithm has previously been used to model infants' language discrimination abilities (Carbajal, Fer, & Dupoux, 2016) and appears to be a good candidate to account for speaker-related effects, since the representations it produces at the utterance-level (so called *i-vectors*) are widely used—among other things—to provide speaker information in the context of speech technology applications (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010). Furthermore, the frame-level part of the learning algorithm has recently been shown to correctly predict language-specific phonetic learning effects observed in infants (Schatz, Feldman, Goldwater, Cao, & Dupoux, 2019). Taken together, this makes the proposed mechanism a good candidate to account for the language familiarity effect in infants.

Methods

Speech recordings

To train and test our models, we use subsets from four corpora of speech recordings, two in American English and two in Japanese, with one corpus of read speech and one corpus of spontaneous speech for each language (Table 1). Read speech in American English is obtained from a corpus of read news articles (Paul & Baker, 1992); spontaneous speech from a corpus of casual conversations (Pitt, Johnson, Hume, Kiesling, & Raymond, 2005). Read speech in Japanese is obtained from a corpus of read news articles (Schultz, 2002); spontaneous speech from a corpus of speakers recounting an event from their life in front of a small audience (Maekawa, 2003). The speech stream is split into separate utterances which are fed to the learning algorithm individually. For each corpus we have a separate training and test set with non-overlapping sets of speakers.

Learning algorithm

To train a model, we start by extracting, for each utterance, moderate-dimensional ($d = 39$) descriptors of the spectral shape of 25 ms-long speech frames sampled every 10 ms along the signal (MFC coefficients (Mermelstein, 1976)), which can be interpreted in terms of auditory pre-processing (Schatz, 2016). The proposed learning mechanism then consists of fitting a hierarchical probabilistic generative model to these descriptors. Specifically, the spectral shape descriptors are assumed to be generated by a mixture of $K = 2048$ (full covariance) Gaussians. Unlike in a plain Gaussian mixture model however, where the mean of each Gaussian is fixed, our model's Gaussian means are assumed to be generated separately for each utterance. The mean vector for Gaussian component k in utterance j is generated by adding an utterance-independent mean vector μ_k to the product of an utterance-independent 39×400 matrix T_k—which encodes the directions in which the mean of the k-th Gaussian component is susceptible to vary from one utterance to the next—and a relatively low-dimensional ($d = 400$) *i-vector* w_j—which encodes characteristics that are specific to a particular utterance and is shared across all Gaussian components, allowing for correlations in the utterance-level displacements of their means. It is these utterance-specific i-vectors that we will use as the infant's representation of an utterance when simulating speaker discrimination tasks. The full generative model is represented in Figure 1, where the depicted variables have the following conditional distributions:

$$\begin{array}{lcllcl} z_{ij} & \mid & \pi_1, \pi_2, ..., \pi_K & \sim & \mathrm{Cat}(\pi_1, \pi_2, ..., \pi_K) \\ m_{kj} & \mid & \mu_k, T_k, w_j & \sim & \delta_{\mu_k + T_k w_j} \\ X_{ij} & \mid & z_{ij}, (m_{kj})_{k=1}^{K}, (\Sigma_k)_{k=1}^{K} & \sim & \mathcal{N}(m_{z_{ij} j}, \Sigma_{z_{ij}}) \end{array}$$

where $\mathrm{Cat}(\pi_1, \pi_2, ..\pi_K)$ is the categorical distribution, π_k is the probability of a point being generated from Gaussian k, δ_x is the dirac delta function with unit probability mass at x and $\mathcal{N}(\mu, \Sigma)$ is the multivariate normal distribution with mean μ and covariance Σ.

We train the model using expectation maximization (Dempster, Laird, & Rubin, 1977) to find parameters that assign

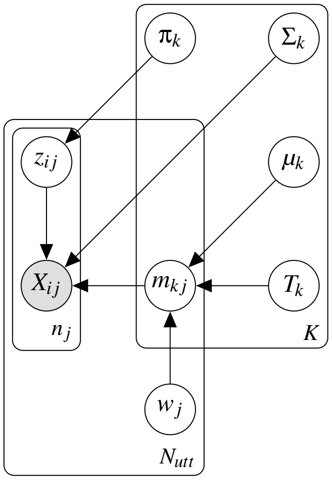

Figure 1: Graphical representation of the generative model.

a high likelihood to the training data. Parameters for a plain diagonal-covariance Gaussian mixture are first fit and used as a seed to train a plain full-covariance Gaussian mixture. The parameters of the mixture are then frozen and the T_k vectors characterising the utterance-level variability in the Gaussian means are fit, also through expectation maximization. This was implemented by adapting the sre08 recipe from the Kaldi speech recognition toolkit (Povey et al., 2011).

Simulating a speaker discrimination task

We compare the ability of our models to discriminate between speakers in their 'native' language and in a 'foreign' language using the machine ABX paradigm (Schatz et al., 2013; Schatz, 2016), which has been used in the past to obtain predictions from computational models and compare them with human behavior (Schatz, Bach, & Dupoux, 2018; Schatz et al., 2019). A detailed computational model of infants' performance in speaker discrimination tasks would be hard to constrain given the limited amount of data available from infant experiments, and is not likely to be necessary, because different discrimination paradigm are known to lead to correlated experimental results (Macmillan & Douglas, 2005). The machine ABX task has the advantage of being a simple, effectively parameterless, evaluation procedure, whose results can reasonably be expected to be qualitatively similar to what would be obtained with more detailed models.

We implement a discrimination task in which the model is presented with two utterances from one speaker - A and X - and a third utterance from a different speaker - B, and has to determine whether X is closer to A or B on the basis of its representations for these three utterances. As the model representation for an utterance, we use the i-vector inferred by the model for that utterance, which provides a fixed-dimensional representation of its global acoustic characteristics. The model 'answers' that X is closer to A than to B if the Euclidean distance $||i_A - i_X||_2$ between the model's i-vector representations of A and X is lower than the Euclidean distance $||i_B - i_X||_2$ between the model's i-vector representations of B and X. If X is judged to be closer to B, the model makes an error, and we define the *ABX error rate* as the probability of an error in this task for two speakers selected at random among the test set's speakers and utterances A, B, X selected at random from the test set utterances available for those speakers. An ABX error rate of zero indicates perfect discrimination and 0.5 indicates chance performance.

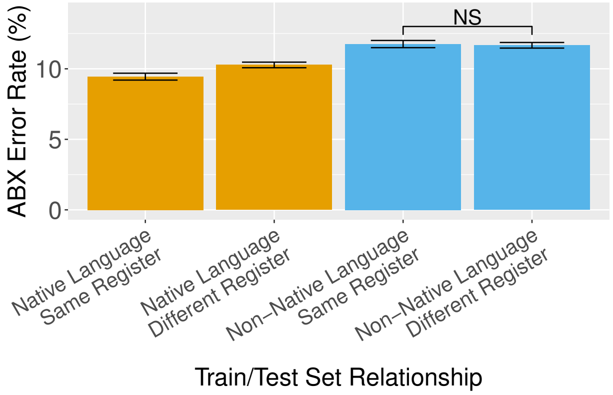

Figure 2: Machine ABX error rate in a speaker discrimination task as a function of the relationship between training and test set languages and registers. Orange corresponds to speaker discrimination in the 'native language', blue to speaker discrimination in a 'foreign language'. Speakers are easier to discriminate in the 'native language' conditions compared to the 'foreign language' ones, irrespective of whether registers match or not, which constitutes strong evidence for a language familiarity effect. Effects of register mismatch appear small overall. Error bars represent plus and minus one standard error of the mean and were obtained through a minimum variance unbiased estimator taking into account the dependencies due to the recurrence of a same speaker in multiple speaker pairs (Lee, 1990). All pairwise comparisons were highly significant after Bonferroni correction for multiple comparisons in asymptotic z-tests on paired difference scores ($p < 3 \times 10^{-11}$), except for the one labeled NS ($p = 0.53$).

Results

Separate models were trained on the training set of each of the four corpora. Each model was then tested with utterances from the test set of each corpus, meaning that each model was tested on both read and spontaneous speech in its native language and a foreign language. We thus obtained a total of 16 ABX error rates (4 models times 4 test sets), which we binned into 4 conditions according to whether the training and test sets' language and register matched or not. The models show a language familiarity effect, i.e. they are better at discriminating speakers when tested in their 'native' language, as evidenced by a lower rate of ABX discrimination errors in conditions where the training language and test language are the same (Figure 2). This is the case even when the training and test register are different, showing the robustness of the language familiarity effect found in these simulations.

Discussion

We proposed a learning mechanism for the language familiarity effect in infants and demonstrated its feasibility. Our results show that sophisticated phonological knowledge and language understanding are not necessary prerequisites for a language familiarity effect to emerge. Instead, we can capture the language familiarity effect simply by modeling acoustic variability in the signal at frame- and utterance-level time scales.

One outstanding question is why the language familiarity effect observed in adults seems to require sophisticated linguistic knowledge, whereas we find that an effect can emerge directly from modeling acoustics. Fecher et al. (2019) proposed that this discrepancy might arise from the different experimental tasks used in infants and adults. Whereas the effect in infants has been shown in speaker *discrimination* tasks, speaker *identification* tasks have typically been used with adults. Identification tasks tap into more sophisticated cognitive abilities, which might explain why they lead to an effect that relies on abstract linguistic knowledge. Follow-up work should investigate whether adults exhibit a language familiarity effect that relies on purely acoustic factors in discrimination paradigms.

Our approach consists of explicit simulation of learning occurring in infants in ecological situations outside of the lab, followed by probing of the learned models in ways that mirror the experimental assessment of infants' abilities in the lab. One important benefit of this approach—compared to the more typical approach consisting of fitting models directly to experimental data from laboratory experiments—is that the exact same model can be evaluated in multiple experimental tasks, just like the same experimental participant can perform multiple tests. Indeed, there is now evidence that the learning mechanism we propose can simultaneously account for language familiarity, language discrimination (Carbajal et al., 2016) and phonetic learning (Schatz et al., 2019) effects observed in infants. We believe that focusing on modeling ecological cognitive functions that can account for observed behavior in multiple experimental settings is going to be instrumental in the quest for a more unified understanding of intelligent behavior.

Acknowledgements This research was supported by NSF grant BCS-1734245 and ESRC grant ES/R006660/1. We thank Elizabeth Johnson for helpful discussion.

References

Bergelson, E., & Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 109, p. 3253-3258).

Carbajal, M. J., Fer, R., & Dupoux, E. (2016). Modeling language discrimination in infants using i-vector representations. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 889–894).

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

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

Fecher, N., & Johnson, E. K. (2018). The native-language benefit for talker identification is robust in 7.5-month-old infants. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(12), 1911–1920.

Fecher, N., & Johnson, E. K. (in press). By 4.5 months, linguistic experience already affects infants talker processing abilities. *Child Development*.

Fecher, N., Paquette-Smith, M., & Johnson, E. K. (2019). Resolving the (apparent) talker recognition paradox in developmental speech perception. *Infancy*, *24*(4), 1–19.

Gervain, J., & Mehler, J. (2010). Speech perception and language acquisition in the first year of life. *Annual review of psychology*, *61*, 191–218.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*(5), 448–458.

Johnson, E. K., Bruggeman, L., & Cutler, A. (2018). Abstraction and the (Misnamed) Language Familiarity Effect. *Cognitive Science*, *42*(2), 633–645.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011.

Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics. The International Journal of Speech, Language and the Law*, *4*(1), 18–28.

Lee, A. J. (1990). *U-statistics: theory and practice.* Marcel Dekker.

Macmillan, N. A., & Douglas, C. C. (2005). *Detection theory: A user's guide, 2nd ed.* Associates Publishers.

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

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, *116*, 91-103.

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).

Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*, 1899–1910.

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.*

Schatz, T. (2016). *ABX-discriminability measures and applications* Doctoral dissertation. Université Paris 6 (UPMC).

Schatz, T., Bach, F., & Dupoux, E. (2018). Evaluating automatic speech recognition systems as quantitative models of cross-lingual phonetic category perception. *The Journal of the Acoustical Society of America*, *143*(5), EL372–EL378.

Schatz, T., Feldman, N., Goldwater, S., Cao, X. N., & Dupoux, E. (2019). *Early phonetic learning without phonetic categories – Insights from machine learning* (preprint). PsyArXiv.

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.*

Schiller, N., Koster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics. The International Journal of Speech, Language and the Law, (1997)*, *4*(1), 1–9.

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