Title

Modeling language discrimination in infants using i-vector representations

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 38(0)

Authors

Carbajal, M. Julia
F´er, Radek
Dupoux, Emmanuel

Publication Date

2016

Peer reviewed

Modeling language discrimination in infants using i-vector representations

M. Julia Carbajal [1] **(carbajal.mjulia@gmail.com)**
Radek Fér [2] **(ifer@fit.vutbr.cz)**
Emmanuel Dupoux [1] **(emmanuel.dupoux@gmail.com)**

[1]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS/EHESS/CNRS; 29, rue d'Ulm
75005 Paris, France
[2]Speech@FIT, Faculty of Information Technology, BUT; Božetěchova 2
612 66 Brno, Czech Republic

Abstract

Experimental research suggests that at birth infants can discriminate two languages if they belong to different rhythmic classes, and by 4 months of age they can discriminate two languages within the same class provided they have been previously exposed to at least one of them. In this paper, we present a novel application of speech technology tools to model language discrimination, which may help to understand how infants achieve high performance on this task. By combining a Gaussian Mixture Model of the acoustic space and low-dimensional representations of novel utterances with a model of a habituation paradigm, we show that brief exposure to French does not allow to discriminate between two previously unheard languages with similar phonological properties, but facilitates discrimination of two phonologically distant languages. The implications of these findings are discussed.

Keywords: language discrimination; speech; acoustics; computational models; habituation

Introduction

When infants acquire their first language, they meet the formidable challenge of dealing with massive variability and ambiguity at all levels of acoustic and linguistic structure. Infants growing up in a multilingual environment must face an additional level of variability due to the presence of two (or more) languages with independent yet partially overlapping acoustic and structural properties. Although the task may seem hard, a large number of studies show that the ability to discriminate spoken languages is present early on in life (Mehler et al., 1988; Nazzi, Bertoncini, & Mehler, 1998; Nazzi, Jusczyk, & Johnson, 2000; Bosch & Sebastian-Galles, 2001; Ramus, 2002; Byers-Heinlein, Burns, & Werker, 2010). For example, using a habituation paradigm, Mehler et al. (1988) showed that French newborns, in spite of their brief experience with language, are able to discriminate their native language from a foreign one (in this case, Russian) as evidenced by an increase in their arousal following a switch from Russian to French utterances. This discrimination was still observed when infants were presented with low-pass filtered speech, and a preference for their native language was suggested by an asymmetry in the arousal depending on the language presented during habituation.

Further research extended these findings, supporting the claim that newborns can distinguish any two unheard languages if they belong to different rhythmic classes, such as Japanese and English, but that they fail to do so if they belong to the same rhythmic class, e.g., English and Dutch (Nazzi et al., 1998; Byers-Heinlein et al., 2010). These results point at prosody as a strong cue for language discrimination at an early developmental stage. However, languages often differ in many other dimensions, such as their phonemic inventories and phonotactic rules. These cues may become relevant through further exposure to one or more languages and thus facilitate their discrimination: by 4 to 5 months of age, both monolingual and bilingual infants can discriminate two languages even within the same rhythmic class, such as Spanish and Catalan, if they were exposed to at least one of them before (Nazzi et al., 2000; Bosch & Sebastian-Galles, 2001).

While these studies suggest that language distance plays an important role, the specific acoustic features and mechanisms that may allow language discrimination throughout the first year of life, and the impact of prior exposure to one or more languages, are not yet fully understood. In the present study we explore how state-of-the-art speech technology tools can help us understand this feat. As a first step in the application of these novel techniques to the study of infant perception, we propose the use of *i-vectors*, a method to represent any given utterance as a pattern of deviations from a previously constructed background acoustic distribution, to implement an unsupervised model of language discrimination. The i-vector representation, in combination with discriminative classifiers, was originally developed for automatic Speaker Recognition (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010), and in recent years has been adapted to Language Identification systems showing excellent performance (Martínez, Plchot, Burget, Glembek, & Matějka, 2011). These models are typically trained on large datasets containing many different speakers/languages to capture all possible sources of variability. Here, we simplify the model to represent the brief experience of an infant exposed to a single speaker of French, and then test the system's ability to discriminate new unheard utterances of two languages that differ in many phonological dimensions, such as rhythm, syllabic structure and phonemic repertoire (French and English), and two languages with largely overlapping phonologies (Spanish and Catalan). As most studies of language discrimination have made use of habituation paradigms, we also propose a computational model of the habituation task, which will allow us to compare the performance of our system with what has been observed in young infants.

The remainder of the paper unfolds as follows. We first introduce the concept of Universal Background Model and i-vector representation, discussing how these representations can be adapted to model infants' experience. Next, we describe the datasets that we selected for the modeling of the background space and the language discrimination tests. Then, we present a model of the habituation task that uses the extracted i-vectors as input, and two additional measures of discriminability. Finally, the results are described and discussed with respect to current experimental data, followed by a perspective on future work.

Methods

Universal Background Model and i-vectors

The first step of the modeling consists in constructing a representation of the acoustic space formed through the infant's exposure to a given linguistic environment, i.e., their "native" language. To model the distribution of speech features, speech technologies typically use Gaussian Mixture Models (GMM). With a sufficient number of mixture components, GMMs can model any arbitrarily complex distribution. The typical number of components for a Language Identification (LID) system is around one thousand.

The parameters (weights, means and covariances) of the model can be estimated by Maximum Likelihood using an Expectation-Maximization algorithm (Bishop, 2006). A GMM trained on a large database of several hundred hours of speech containing many different speakers, languages and other sources of variability, can be used to represent the overall feature distribution. In the context of speaker and language recognition, this is called the *Universal Background Model* (UBM). Evidently, young infants cannot count on such a large and variable amount of data to build their representations of the acoustic space, however, nothing prevents UBMs from being trained on a much smaller dataset. In the present study, we train a small UBM with speech from one single French speaker to represent the brief exposure that even a 4-day-old infant may have already encountered.

Once the UBM has been trained, data-specific models representing feature distributions of different utterances can be derived from the UBM by Maximum a Posteriori (MAP) adaptations. Usually, only the component means are shifted during the adaptation. Using factor analysis, the adaptation offset with respect to the UBM can be confined to a low-dimensional subspace, called the Total Variability space. If we denote by $\mathbf{m}$ the stacked vector of UBM component means, the generative subspace model has the form:

$$\boldsymbol{\mu} = \mathbf{m} + \mathbf{T}\mathbf{v},$$

where $\mathbf{T}$ is a low-rank matrix (Total Variability matrix) defining the bases for the subspace, and $\mathbf{v}$ is a hidden variable with standard normal prior. As with the UBM, this subspace is typically trained on a large number of speech recordings using EM algorithms (Dehak et al., 2010), but for the purpose of our model it will be trained on the data of a single speaker.

Finally, given an utterance or any other segment of a speech recording, the posterior distribution of the hidden variable can be estimated. The MAP point estimate of this distribution is conventionally called an *i-vector*, and can be used as a low-dimensional fixed-length representation of the speech segment. In other words, any unheard utterance can be approximated as a deviation from the background "native" model. We propose to use this simple representation to model the infant's acoustic perception of previously unheard speech, computing an i-vector for every utterance in our test dataset. The advantage of this vectorial representation of speech is that a measure of distance can be defined between any two utterances.

In LID systems, the typical dimensionality of the subspace is around 400. However, for our experiments, the i-vector dimensionality is set to 200, and we use a UBM with 256 mixture components and diagonal covariance matrices. The reason for such a small model is that the database we propose to use in order to model a brief exposure to French is not large enough to robustly estimate all the parameters of a conventional LID model. Furthermore, since our database contains only a limited amount of variability (UBM trained on one single speaker and language), it is unnecessary to increase the number of dimensions.

We argue that i-vectors are reasonable as models of infants' representation of languages for the following reasons: (1) The entire pipeline (construction of UBM and i-vector extraction) only requires two skills, which have been documented in infants: a good acoustic perception (Eimas, Siqueland, Jusczyk, & Vigorito, 1971), and the ability of performing statistical learning (Saffran, Aslin, & Neport, 1996; Maye, Werker, & Gerken, 2002). (2) The learning algorithm is completely unsupervised, requiring no external information about phonemes or words, nor any information about speaker identity, or number and properties of different languages. The only linguistic hypotheses of this model are that utterances are relevant units for performing language discrimination, that they can be modelled through gaussian mixtures, and that they can be segmented out of continuous speech.

Feature extraction A common representation of the acoustic features of a speech signal used in many speaker and language identification systems are *Mel-Frequency Cepstral Coefficients* (MFCCs), which are based on a transform of the power spectrum on a frequency scale that approximates human auditory perception. For our modeling purposes, these features were calculated using the HTK Speech Recognition Toolkit (Young et al., 2006) in 25 ms windows with a 10 ms shift. We retained the first 7 coefficients (including *C0*, which represents the energy) and added a measure of *F0* (pitch) computed with the Kaldi Toolkit (Povey et al., 2011).

In addition, *Shifted Delta Coefficients* (SDC, a stacked version of delta coefficients calculated across several frames, Torres-Carrasquillo et al., 2002) were included to capture the temporal evolution of the MFCC-F0 features. The SDCs were calculated using the usual 7-1-3-7 configuration, resulting in

an approximation of the contour of the MFCC-F0 features over a span of 200 ms. The resulting 64-dimensional MFCC-F0-SDC vectors contain both spectral and prosodic information presumed available to the human auditory system.

Materials

Training data In order to train the UBM to represent the prior experience of an infant with a brief exposure to French, we used a dataset of casual speech recorded from an adult female French speaker selected from the *Corpus of Interactional Data* (Bertrand et al., 2008). The selected dataset is composed of 602 pre-segmented utterances with a mean length of 2.54 seconds ($min = 0.43$ s, $max = 9.01$ s), giving a total of approximately 25 minutes of clean speech. The original recordings were downsampled to 16kHz.

Test data Similarly to previous experimental studies, to test the discrimination of languages we recorded two proficient bilingual speakers: a male French-English bilingual speaker and a female Spanish-Catalan bilingual speaker. The use of bilingual speakers for the test data aims at reducing any sources of variability not due to the target languages. During each recording session, the speakers read the first two chapters of the book *The Little Prince* in one of their languages, and immediately afterwards they were asked to discuss what they had read. This procedure was then repeated for their second language. All recordings for each speaker were done on a single session.

The audio recordings were semi-automatically segmented into utterances with a 300 ms silence threshold using the speech analysis software *Praat* (Boersma & Weenink, 2014), and subsequently downsampled to 16kHz. The resulting dataset is composed of 319 utterances (French: 65, English: 75, Spanish: 99, Catalan: 80), with a mean length of 3.69 seconds ($min = 2.00$ s, $max = 10.63$ s).

Model of habituation task

Experimental studies of language discrimination in infants use an habituation paradigm (Mehler et al., 1988; Nazzi et al., 1998). In this paradigm, infants are presented with a set of stimuli from one language (L1), and their arousal is measured (in newborns, it is measured with a pacifier connected to a pressure detector). After an initial increase, infants' arousal decreases, indicating habituation. When a threshold has been reached, half of the infants continue with the same class of stimuli, and the other half are switched to a second class (L2). The difference of arousal after the switch in the two groups is used as a measure of discrimination.

Here, we will model this paradigm using an on-line clustering algorithm. In the habituation phase, the system gradually incorporates data from one language (L1) until it reaches a statistical threshold. In the test phase, as for infants, new utterances of L1 (*same* condition) and L2 (*switch* condition) are compared to the habituated model. The input of this model consists of the i-vectors of the test utterances as extracted by our previously trained system. To reduce spurious effects caused by specific subsets of utterances, the habituation task was run 100 times for each language pair using randomly selected subsets in each trial.

Habituation phase The model starts with an initial set of 10 i-vectors $\{v_1, ..., v_{10}\}$ of one language (L1) chosen randomly from our dataset. Firstly, the centroid μ_1 of this initial set (i.e., the mean i-vector) is computed, and the cosine distance of each of the 10 composing vectors to the centroid $d_c(v_i, \mu_1)$ is calculated. Secondly, a new random set of 10 i-vectors $\{v_{11}, ..., v_{20}\}$ of the same language L1 is selected, and their cosine distances to the initial centroid μ_1 are calculated. The distribution of distances of the initial and the second set of vectors are then compared with a t-test.

If $p \leq 0.05$, the two distributions are considered statistically different, that is, the model perceives a difference between the two sets of utterances, and therefore has not yet reached habituation. In this case, the last set of vectors is aggregated to the initial set and the centroid is recalculated, μ_2, as the mean i-vector of the whole set. Following the same procedure, a new group of 10 i-vectors from L1 is selected and their cosine distance to the new centroid $d_c(v_i, \mu_2)$, $\{i = 21, ..., 30\}$, are calculated and compared through a t-test to the distance of the previous 10 vectors to the new centroid $d_c(v_i, \mu_2)$, $\{i = 11, ..., 20\}$. This procedure is repeated as long as $p \leq 0.05$.

When $p > 0.05$ (defined as our saturation threshold), the two distributions are not statistically different and the habituation phase is therefore complete. As a final step, the last group of vectors is aggregated to the previous set and a final centroid is obtained, μ_F. The distance of the last 10 vectors to μ_F is then calculated and retained for the test.

Test phase In this stage, a new set of 10 i-vectors v_i is randomly selected from the same language used in habituation (L1, *same* condition) as well as 10 i-vectors u_j from the second language of the same bilingual speaker (L2, *switch* condition). For each set of vectors, the cosine distance to μ_F is calculated.

We finally perform two t-tests, one per condition, comparing the distribution of distances of the new vectors of L1 or L2 to the distribution of the last 10 habituation vectors. In the *same* condition, as the new utterances belong to the same language as those in habituation, the p-value of the t-test is expected to remain above the saturation threshold, $p > 0.05$. On the other hand, in the *switch* condition, the p-value will depend on the overlap between the distribution of the habituation (L1) and L2: a p-value below the 0.05 threshold would mean that the two distributions are significantly different, indicating discrimination of the two languages, while $p > 0.05$ would indicate a lack of discrimination.

Discriminability measures

To quantify the discriminability of the languages independently of our habituation-dishabituation model, we computed the pairwise ABX discrimination score, a nonparamet-

ric measure of category overlap. It consists in taking all possible ABX triplets of utterances from a language pair, where A corresponds to an utterance of L1, B corresponds to an utterance of L2, and X can be either L1 or L2. For each triplet, X is classified as belonging to L1 or L2 based on whether the cosine distance between X and A is smaller or greater than the distance between X and B. The percentage of correct classifications serves as an index of the discriminability between the two languages. Additionally, we performed a *Principal Component Analysis* (PCA) for each language pair as a way of visualizing the variance and distance of the i-vectors that compose each language.

Results

Habituation task

We ran the habituation model for both language pairs, and within each pair we tested the model with both possible languages in the initial habituation phase. The average amount of steps to reach habituation was similar for all languages (French: 2.1, English: 1.8, Spanish: 1.7, Catalan: 1.7).

As previously observed in infant experiments, the results of 100 trials for each test (presented in Figure 1) show a difference in the pattern of discrimination of the two language pairs. In the case of Spanish-Catalan (bottom panels), the p-values of both the *same* condition and the *switch* condition are significantly above the threshold value of $p = 0.05$, independently of the language presented in habituation (Habituation:Spanish -bottom right panel- *same*: $M = 0.48$, $SD = 0.26$, *switch*: $M = 0.40$, $SD = 0.26$; Habituation:Catalan -bottom left panel- *same*: $M = 0.52$, $SD = 0.28$, *switch*: $M = 0.54$, $SD = 0.27$), suggesting a lack of discrimination of these two languages. On the other hand, the second language pair (French-English, top panels) presented an asymmetry in the responses of the model to the *switch* condition, depending on the language of habituation. When the system is habituated to English as L1 and then switches to French (top left panel), the two languages are discriminated as indicated by a decrease of the p-value below the threshold in the *switch* condition (*same*: $M = 0.49$, $SD = 0.29$, *switch*: $M = 0.012$, $SD = 0.026$). However, if the system is initially habituated to French (top right panel), the switch to English is not detected, with both conditions showing similar p-values (*same*: $M = 0.54$, $SD = 0.29$, *switch*: $M = 0.48$, $SD = 0.25$). While a similar behavior was observed in infant habituation experiments (Mehler et al., 1988), additional analyses are required to understand this asymmetry.

ABX and Principal Component Analysis

To further explore the different response patterns of our model, we performed an ABX task for both language pairs and all possible X categories (ABA, ABB). The results of this test, shown in Table 1, present a similar pattern to the one observed in the habituation task. In the case of Spanish-Catalan, both ABA and ABB trials presented scores slightly above chance level (50%), meaning that nearly half of the Spanish

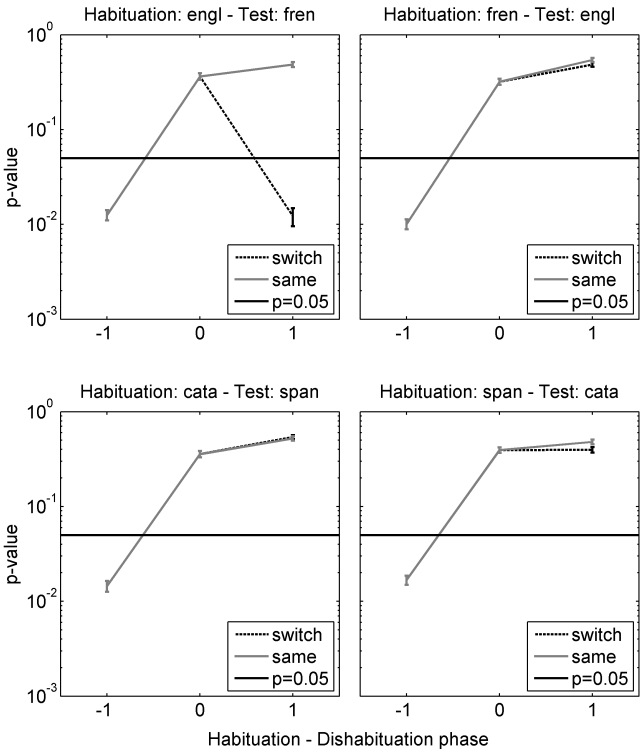

Figure 1: Average p-values over 100 trials of the habituation task for French-English discrimination (top) and Spanish-Catalan discrimination (bottom). The *x* axis represents the steps of the habituation and test phase, where *0* indicates the step where the habituation threshold ($p = 0.05$) was reached. Accordingly, *step -1* represents one step before habituation, and *step 1* represents the test (dishabituation) phase.

utterances were incorrectly categorized as Catalan utterances (and vice-versa). On the other hand, French-English trials presented an asymmetry: a majority of English utterances were correctly classified, while the classification of French utterances remained near chance level. This means that the distance between two given French utterances in the test set is often larger than the distance between a French and an English utterance, pointing at a possible imbalance in the variance of the distributions of their i-vector representations.

Table 1: Summary of ABX results: Percentage of accuracy for the distant language pair (A = English, B = French) and the close language pair (A = Catalan, B = Spanish).

Language Pair	X=A	X=B
English (A) - French (B)	76%	46%
Catalan (A) - Spanish (B)	51%	57%

Finally, we performed a Principal Component Analysis on both language pairs in order to visualize the distribution of the utterances. A representation of the first two dimensions

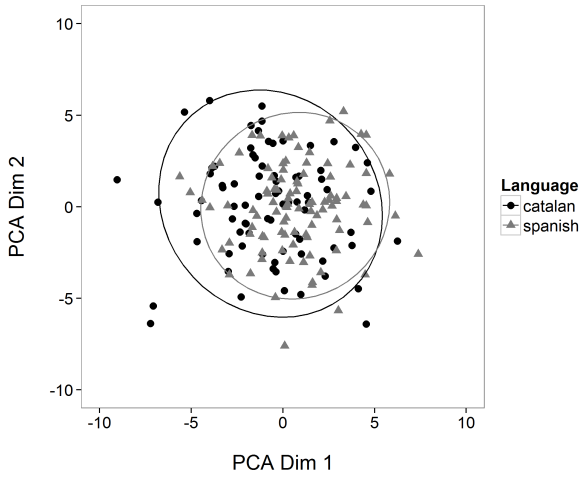

Figure 2: First two dimensions obtained through a Principal Component Analysis of the i-vectors of Spanish and Catalan utterances spoken by a bilingual speaker.

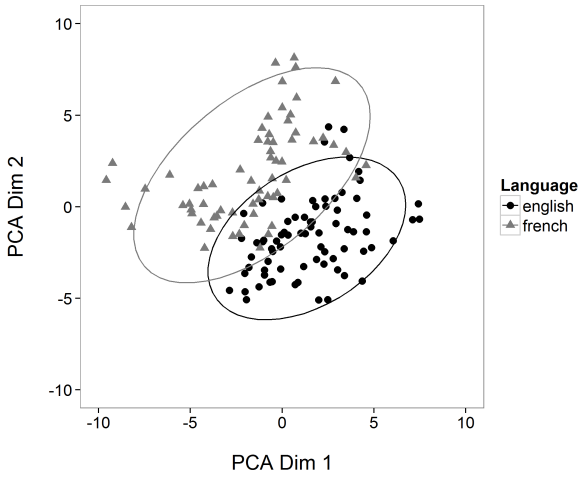

Figure 3: First two dimensions obtained through a Principal Component Analysis of the i-vectors of French and English utterances spoken by a bilingual speaker.

of the PCA for the Spanish-Catalan pair, shown in Figure 2, revealed a high degree of overlap in the distribution of the utterances of these two languages. On the contrary, the first two dimensions of the French-English PCA, presented in Figure 3, show a higher separation between the two languages. However, as suggested by the ABX score, the variance in these two dimensions appears to be larger within French utterances than within English utterances.

Together with the ABX results, this difference in the variance may explain the asymmetry observed in the habituation task: when the model is habituated to English, the variance of the i-vectors that are aggregated during this initial phase remains small, allowing the system to detect a switch to the second language. In other words, the within-language distance distribution is smaller than across-language. However in the inverse case, when the model is initialized with French, the variance of the habituation vectors is relatively large and therefore the switch to English remains unnoticed.

In summary, we found an overall difference in the degree of separation of the i-vectors of both language pairs, which reflected in the behavior of our habituation-dishabituation model. Spanish-Catalan utterances present largely overlapping distributions, causing a lack of discrimination in the habituation task, while French-English utterances have less overlapping yet more asymmetrical distributions, producing an equally asymmetric response of the system.

Discussion

In this paper we introduced a novel application of speech technology tools to model language discrimination in infants. Using a GMM-UBM trained on a small dataset of French utterances, we represented the acoustic space of a monolingual infant after a brief exposure to this language. To test the system's ability to discriminate languages, we modelled the acoustic representation of novel utterances as a pattern of shifts from the means of the UBM. Using this low-dimensional representation, called i-vector, we constructed a model of a habituation task similar to the experimental paradigm often used to test discrimination in infants.

The behavior of our model in the habituation task resembled that observed in previous experiments: our system, pre-exposed to French, was unable to discriminate between two previously unheard languages with highly similar phonologies (Spanish & Catalan), while it discriminated two phonologically distant languages (French & English). Interestingly, just as reported in previous infant studies such as Mehler et al. (1988), the ability to discriminate between French and English depended on the language presented during habituation. When the system was initially habituated to the previously unheard language (English), it detected a switch to the "native" language (French), but it failed at discriminating a switch to English when French was presented in habituation. Further analyses provided a potential explanation for our model's asymmetrical behavior: the variance of the i-vector representations of French utterances is larger than that of English utterances, causing the habituation model to create a broad category for French which hinders the discrimination of English. While in the context of infant studies this asymmetry was regarded as a preference for the native language, our modeling results suggest that the perceived acoustic variability might be responsible for this behavior, providing a new perspective on this issue.

There are three possible explanations for the larger variance of French as compared to English in our test data. First of all, this difference might be a characteristic of the specific bilingual speaker that was recorded for this experiment. To test this hypothesis, it would be necessary to repeat the test with a different French-English speaker. If the same pattern

was observed, it would indicate that the difference does not lie in the speaker but in the language. This could mean that, overall, French speech is acoustically more variable than English. However, and more interestingly, it is also possible that the difference was originated in the training of the Universal Background Model and the Total Variability subspace: as our system was pre-exposed only to French, the model may have developed a larger sensitivity to acoustic differences present in French speech than those in English speech, thus appearing more variable. To discern these two possibilities, the model could be re-trained using English as the background (i.e., "native") language. If the larger variance is due to the sensitivity of the model to its native language, then the asymmetry should be inverted. The results of these future modeling experiments may help us better understand the behavior observed in infants.

In addition, this methodology can be applied to model language discrimination in a variety of other cases. First, the UBM and the TV subspace can be trained with different languages and with varying amounts of data to investigate the impact of language exposure on discrimination (e.g., the model can be trained with a large dataset of Spanish speech and then tested on its ability to discriminate Spanish from Catalan). Second, the system could be trained with a bilingual background to study how multilingualism affects the construction of the acoustic space and consequently its ability to discriminate languages. This bilingual background can be composed of either monolingual speakers of two languages or bilingual speakers, giving further insight into the impact of different bilingual environments on the perceptual system. Third, the acoustic features provided to the model can be adapted (for example, by using filtered speech, or adding additional prosodic information to the feature vectors) to explore the role of different cues in language discrimination. The experimental data available to date provides a means of evaluation for the models, which in turn may generate new testable hypotheses that will help us better understand how young infants achieve this task.

Acknowledgements

We thank Alexander Martin and Laia Fibla for their participation in the recordings, and Hynek Heřmanský and Lukáš Burget for their helpful discussions. This work was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON), the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the École des Neurosciences de Paris Ile-de-France, the Region Ile de France (DIM Cerveau et Pensée), and an AWS in Education Research Grant award.

References

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., & Rauzy, S. (2008). Le CID - Corpus of Interactional Data - Annotation et exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, *49*(3), 1–30.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer [Computer program].* Retrieved from `http://www.praat.org` (Version 5.3.86)

Bosch, L., & Sebastian-Galles, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy*, *2*(1), 29–49.

Byers-Heinlein, K., Burns, T. C., & Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological Science*, *21*(3), 343–348.

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

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Martínez, D. G., Plchot, O., Burget, L., Glembek, O., & Matějka, P. (2011). Language recognition in ivectors space. In *Proceedings of interspeech 2011* (Vol. 2011, pp. 861–864). International Speech Communication Association.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology*, *24*(3), 756–766.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by english-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, *43*, 1–19.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* IEEE Signal Processing Society.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, *2*(1), 85–115.

Saffran, J. R., Aslin, R. N., & Neport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., & Deller, J. (2002). Approaches to language identification using Gaussian Mixture Models and Shifted Delta Cepstral features. In *ICSLP 2002* (pp. 89–92).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., ... others (2006). The HTK book (for HTK version 3.4).