Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors

Maureen de Seyssel (maureen.deseyssel@gmail.com)
Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
Laboratoire de Sciences Cognitives et Psycholinguistique, ENS-PSL/EHESS/CNRS/INRIA
Paris, France

Abstract

The language discrimination process in infants has been successfully modeled using i-vector based systems, with results replicating several experimental findings. Still, recent work found intriguing results regarding the difference between monolingual and mixed-language exposure on language discrimination tasks. We use two carefully designed datasets, with an additional "bilingual" condition on the i-vector model of language discrimination. Our results do not show any difference in the ability of discriminating languages between the three backgrounds, although we do replicate past observations that distant languages (English-Finnish) are easier to discriminate than close languages (English-German). We do, however, find a strong effect of background when testing for the ability of the learner to automatically sort sentences in language clusters: bilingual background being generally harder than mixed background (one speaker one language). Other analyses reveal that clustering is dominated by speakers information rather than by languages.

Keywords: language discrimination; language diarization; i-vectors; bilingualism; speaker information

Introduction

Bilingualism is a widespread phenomenon, with the majority of children being born in a bilingual environment. It also appears that being raised bilingual does not result in any particular delay in the language acquisition milestones of children compared to the monolingual peers (Oller, Eilers, Urbano, & Cobo-Lewis, 1997; Vihman, Thierry, Lum, Keren-Portnoy, & Martin, 2007; Petitto et al., 2001), nor to any confusion between the different languages (Petitto & Holowka, 2002; Byers-Heinlein & Lew-Williams, 2013). In fact, infants from both monolingual and bilingual environments seem to be able to discriminate between distant languages from birth (Byers-Heinlein, Burns, & Werker, 2010; Mehler et al., 1988), and rhythmically similar languages as young as 5 months old (Nazzi, Jusczyk, & Johnson, 2000; Bosch & Sebastián-Gallés, 1997). How do they do it? What kind of computational system can achieve language discrimination from the raw signal only? Are there pairs of languages or language backgrounds which would make such discrimination easier or harder? One way of addressing these questions is to use automatic language discrimination techniques as a model of how infants process and discriminate languages.

Related work

I-vectors (Dehak, Torres-Carrasquillo, Reynolds, & Dehak, 2011) are fixed-length vector representations of entire utterances which characterize how much an utterance deviates acoustically from a background distribution of speech used to train the system. These representations are typically used for speaker identification and discrimination (Dehak et al., 2011) but can also represent languages (Martinez, Burget, Ferrer, & Scheffer, 2012; Martinez, Plchot, Burget, Glembek, & Matějka, 2011).

I-vectors based systems have been shown to reproduce key findings in language discrimination experiments: the ability to detect a change in language within a bilingual speaker (language discrimination) (Carbajal, Dawud, Thiollière, & Dupoux, 2016), the distance effect between different language pairs, with close languages being harder to discriminate than more distant languages (Carbajal, 2018), and the ability to discriminate based on prosody (Martinez, Lleida, Ortega, & Miguel, 2013; Carbajal, 2018). However, they also resulted in an intriguing prediction that has not so far been verified experimentally. Notably, Carbajal et al. (2016) found that learners exposed to a mixture of languages have more difficulties to discriminate languages than learners exposed to monolingual backgrounds. These results are counter-intuitive: one would think that having a mixed background should help discrimination not hinder it. They also have potentially important empirical and practical implications. Indeed, if true, they would reveal an undocumented discrimination deficit for infants in a bilingual or mixed background. This is why we wanted to replicate them with more controlled stimuli. Indeed, the initial study used English and Xitsonga recordings from completely different datasets, raising the possibility that results might come from recording-specific properties rather than the language characteristics.

Present work

The mixed background deficit effect found by Carbajal et al. (2016), if true, is important both for theoretical and practical reasons. The current study is devoted to reproducing the original effect, test its robustness, and to more fully understand how language background may affect a learner's ability to discriminate languages.

The first aim of the study is to reproduce the original experiments using more controlled and ecological stimuli. First, to discard potential acoustic artifacts, all recordings used in the experiment were from the same corpus. Second, we used a better counterbalancing design allowing the different con-

ditions to be perfectly comparable, all containing the exact same recordings. Third, the datasets are also more ecological, containing a smaller number of speakers ($N = 12$), simulating an infant's exposure to speech better than the original study containing an implausible number of speakers ($N = 168$).

We also introduce three novelties to explore the robustness of the results. First, we compare two language pairs, one being closely-related (English and German) and the other one being more distant (English and Finnish). Besides enhancing the generalizability of the results, this also allows us to test whether close language pairs are more difficult to discriminate than distant language pairs. Second, along the monolingual and mixed conditions, we introduce a new "bilingual" background condition, with speech from the same speakers speaking in both languages. This new condition simulates an environment in which the infant is exposed to bilingual speech from the same persons (e.g. parents switching constantly between language A and language B). Recent theories in psychology support the idea that such a fully bilingual environment can harm the children's linguistic development and therefore suggests that parents should follow the "One Parent, One Language" (or OPOL) strategy (Genesee, 1989). We are therefore able to investigate whether, in modeling language discrimination, a mixed environment (OPOL) and a fully bilingual environment result in any processing differences. Finally, we analyze the effect of speaker information on language discrimination. This was partly done in Carbajal et al. (2016) by applying a Linear Discriminant Analysis (LDA) to the i-vectors to select a new representation that increases the separation between speaker. Here we add a method which, by taking the orthogonal complement of this LDA representation, does the opposite, i.e. normalizes the representation across speakers.

Finally, to more fully understand how language background could affect discrimination, we test language discrimination in two different ways. The first one is based on psycholinguistic experiments run in infants in the laboratory. In such experiments, infants are presented with sentences from a single bilingual speaker speaking one of their languages, and the reaction of the infant to an unpredictable change in language is measured (through behavioral proxies such as looking time or non-nutritive sucking). Children are said to discriminate the two languages if there is a statistical difference between the set of children who had a switch of language and those who did not. As in (Carbajal et al., 2016), we model this task with a machine-ABX discrimination metric (Schatz et al., 2013). We argue, however, that contrary to the standard interpretation of the discrimination paradigm, a statistical difference between groups is not fully ecological. It does not necessarily indicate that infants can sort out individual utterances from their environment according to their language. In practice, infants are not confronted with a single speaker, but with multiple ones, the decision has to be made sentence by sentence (sometimes words by words in the presence of code switching), and the number of languages that they speak is unknown. This second problem can be defined as a **language diarization task**, which we model as a clustering problem. More precisely, we apply a clustering algorithm to the modeled acoustic space of the different training backgrounds, and look at the extent to which the formed clusters correlate with language labels.

Methods

Materials

We used the EMIME bilingual corpus (Wester, 2010). It is a read speech corpus containing bilingual speech (utterances from two languages recorded by the same speaker) with a 16kHz sampling rate. It was split into two datasets, one with English and Finnish speech, and the other with English and German speech. In each subset, the speakers are bilingual, although English is always their second language. For each language, each speaker reads on average twice the same set of 145 sentences, leading to some sentence repetitions in the train set.

We designed three conditions for each dataset: a *monolingual* one composed of speech from a single language; a *mixed* condition in which the two languages are represented but with each person speaking only one of the two languages; and a *bilingual* condition, containing speech from both languages, uttered by the same speakers. To ensure all conditions are fully comparable, we further split the training sets into subsets. Each subset was used independently, and results were then averaged within the conditions. This way, within each dataset (English-Finnish and English-German), each averaged condition contains the *same speech utterances*. A summary of the different training conditions is presented in Table 1. The average utterance duration is of 4.44 seconds in the English-Finnish dataset and 4.52 seconds in the English-German dataset. The total duration of each training set was therefore between 4h23 and 4h37. Additionally, a test set was created for each dataset, using bilingual speech from the highest-rated accent male and female for each language (2 speakers per set). Each test set is composed of 200 utterances (100 per language).

Pipeline

The following section describes the methodology behind the different steps carried out in the experiment. The whole workflow is applied independently to each training set. Unless stated otherwise, the open-source tool Kaldi (Povey et al., 2011) was used for the different stages of the process.

Feature Extraction Mel frequency cepstral coefficients (MFCCs) features (Mermelstein, 1976) were extracted for all train and test sets, with 13 coefficients (including energy). They were calculated on 25ms speech frames, using 10ms shift. These features, widely popular in speech processing, are based on human perception and are therefore adequate for modeling cognitive processes of speech. Shifted-delta coefficients (SDC) are also calculated. They capture long-distance

Table 1: Summary of train datasets

Dataset	Background	N speakers (N males)	N utterances
English-Finnish	Mono	12 (6)	6910
	English	*6 (3)*	*3480*
	Finnish	*6 (3)*	*3430*
	Bilingual	12 (6)	6910
	subset 1	*6 (3)*	*3454*
	subset 2	*6 (3)*	*3456*
	Mixed	12 (6)	6910
	subset 1	*6 (3)*	*3480*
	subset 2	*6 (3)*	*3430*
English-German	Mono	12 (6)	6960
	English	*6 (3)*	*3480*
	German	*6 (3)*	*3480*
	Bilingual	12 (6)	6960
	subset 1	*6 (3)*	*3504*
	subset 2	*6 (3)*	*3456*
	Mixed	12 (6)	6960
	subset 1	*6 (3)*	*3480*
	subset 2	*6 (3)*	*3480*

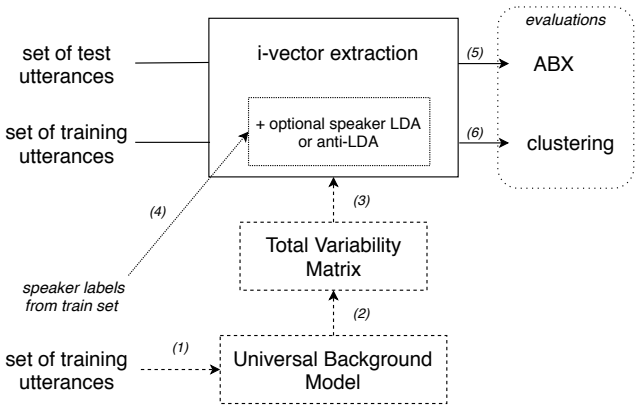

Figure 1: The different stages of the experimental pipeline. In a first training phase, indicated by dotted lines, we construct an i-vector extractor in three steps (1,2,3), followed by an optional step enhancing or reducing the effect of talker variability (4). In the evaluation phrase, indicated by plain lines, we either run a machine equivalent of a discrimination task on novel sentences (5), or cluster the training utterances (6).

information from the neighboring frames, adding some dynamic information to the speech structure.

I-vectors model Following the I-vector model (Dehak et al., 2011), a Gaussian Mixture Model (GMM) is first trained over all speech features of the train set, resulting in a large probabilistic representation of the acoustic space called Universal Background Model (UBM). It can be defined by a supervector m containing the means of all gaussian components. Using factor analysis, the components of highest variability are then projected into a low-dimensional space, the Total Variability space, which is defined by a Total Variability matrix T. An utterance μ can then be defined as $\mu = m + Tv$. The variable v can be used as a fixed dimension representation of μ, and is typically referred to as an i-vector. This process is depicted in Figure 1. We extracted i-vectors for utterances of both the test and train sets. We used a GMM with 128 Gaussians, and dimensionality of 150 for the i-vectors, as these parameters seemed to yield satisfactory results in small datasets (Carbajal et al., 2016).

LDA and Orthogonal Complement Two additional steps were also optionally performed, in an attempt to investigate the effect of speaker information on language discrimination. These supervised methods, applied on the i-vectors, use the speaker labels from the train set to either enhance or diminish the speaker information. They assume that the child is able to identify speakers on an independent basis, and uses this information to either amplify speaker separation or decrease it. To increase speaker information, Linear Discriminant Analysis (LDA) based on the speaker labels is computed on the i-vectors from the train set to estimate a transformation matrix which maximizes the distance between speakers. I-vectors from the train and test sets are then transformed using this matrix, resulting in i-vectors of dimension 11 (*N_speakers* - 1). The opposite stance was also taken by calculating the orthogonal complement of the LDA subspace and then using it to transform the i-vectors. This allowed us to retrieve all the information from the initial i-vector space excluding the information which is in the LDA. By doing this, we remove the information which is used to maximize the distance between speakers, normalizing all speaker information. For clarity reasons, we refer to this extra step as "anti-LDA". The orthogonal complement was calculated using the *scipy* Python package (Virtanen et al., 2019). In cognitive terms, this process would amount to the ability of a child to identify the cues which are speaker-specific, and then removing them from the language identification processes.

Evaluation Methods

Two evaluation methods were implemented, each focusing on one of the language discrimination and language diarization processes.

ABX Scores Language discrimination experiments in psycholinguistics often consist in a first familiarization phase during which the child is exposed to speech from a language A, and an evaluation phase during which the child is presented with two sentences uttered by a new speaker, one of the sentence being from the same language A, and the second sentence being from a novel language B. If the infant can discriminate between the two languages, there should there-

fore be a surprise effect when language B is presented. Although this method is often used as a proxy to assess if children automatically differentiate languages, it is strictly a way to evaluate if children are able to discriminate between two languages, and we therefore restrict our discussion of such results to this particular set-up.

We use the machine ABX paradigm (Schatz et al., 2013) to simulate such a language discrimination experiment. This is done by computing, over the whole set of test i-vectors, multiple triplets of items A, B and X; A and X being i-vectors of utterances sharing the same language and B being an i-vector from an utterance of a different language. For each triplet, the cosine distances of A to X and B to X are then computed. If the distance between A and X is smaller than the distance between B and X, a score of 1.0 is attributed to this triplet, otherwise the score is 0.0. The average of scores across all triplets is then computed, yielding an average ABX score. Perfect discrimination would therefore yield an ABX score of 1.0 (or 100%), as the distance of same-language utterances would always be smaller than the distance of utterances from different languages. To compare our results to psycholinguistics experiment, we compute the triplets within speaker, that is all three items A, B and X will always share the same speaker.

Clustering As a proxy for evaluating whether children cluster multilingual speech from their environment into languages, we apply a clustering algorithm with K clusters to the i-vectors from the multilingual train sets (bilingual and mixed), and evaluate the purity of the formed clusters. If languages are perfectly clustered in the acoustic space, we would expect a purity score of 1.0 when $K = 2$. K-means algorithm was ran 20 times for each K, in the range of $K = 2$ to $K = 20$, yielding an average and standard deviation of the purity scores for each K. We also extended this method to calculate the purity scores on speaker clusters, with $K = 12$ (i.e. the accurate number of speakers). This method was applied to the raw i-vectors from the train sets, as well as the LDA and anti-LDA transformed i-vectors.

Results

ABX scores / Language discrimination

Within speakers ABX scores were computed on the raw, LDA and anti-LDA test i-vectors for each train condition. Results for each dataset are presented in Table 2. Scores in both datasets suggest that the i-vectors successfully allow discrimination between the two languages in all conditions and datasets (no discrimination would yield chance level scores at 50%). As expected, scores in the English-Finnish (different language family) dataset are significantly higher than those in the English-German (same language family) dataset.

There does not seem to be any significant difference with the raw i-vectors between the bilingual, mixed and monolingual conditions, suggesting that the input type in the background's composition does not have an effect on language discrimination of unknown speech. Removing speaker information from the test i-vectors (using the anti-LDA transformation matrix estimated on the train i-vectors) very slightly lowers the discrimination scores in all conditions. In both datasets, however, enhancing speaker information with LDA leads to an increase in ABX scores in the mixed condition, which can be explained by the additional use of speaker information in the language discrimination task, each speaker only corresponding to a single language. It does not yield a stable pattern for the monolingual and bilingual conditions, deteriorating the scores in the English-Finnish dataset but leading to a slight increase in the English-German dataset.

Table 2: Summary of ABX results (in % correct) in both datasets for the different training backgrounds, on the standards, LDA (+LDA) and anti-LDA (-LDA) i-vectors. The scores are calculated within speaker.

		ABX scores		
Dataset	Background	standard	+ LDA	- LDA
English-Finnish	Bilingual	75.1	66.0	74.4
	subset 1	*73.1*	*67.0*	*72.3*
	subset 2	*77.1*	*65.0*	*76.5*
	Mixed	75.5	88.7	73.2
	subset 1	*76.4*	*91.1*	*74.1*
	subset 2	*74.6*	*86.2*	*72.2*
	Mono	73.7	68.0	72.6
	English	*71.8*	*68.9*	*70.4*
	Finnish	*75.5*	*67.0*	*74.8*
English-German	Bilingual	63.3	65.3	62.8
	subset 1	*62.5*	*61.8*	*61.9*
	subset 2	*64.0*	*68.8*	*63.7*
	Mixed	64.2	72.5	62.6
	subset 1	*63.4*	*77.1*	*61.7*
	subset 2	*64.9*	*67.8*	*63.4*
	Mono	63.6	64.9	62.9
	English	*63.1*	*63.3*	*62.5*
	German	*64.1*	*66.4*	*63.2*

Clustering / Language diarization

Kmeans clustering with K clusters (from $K = 2$ to $K = 20$) was applied to the train i-vectors in each mixed and bilingual conditions. Results are presented in Figure 2. The purity score for $K = 2$ is close to 0 in all conditions with, for the raw i-vectors, an average of 0.090 (S = .083) in the mixed condition and of 0.003 (S = .008) in the bilingual condition. This suggests that the acoustic space is not clustered primarily by language.

As presented in Figure 2, with the raw i-vectors, the larger the number of clusters, the larger the difference between the mixed and bilingual conditions, with clusters in the mixed condition getting significantly higher language homogeneity scores. In the mixed condition, language identity is fully correlated with speaker identity, whereas there is absolutely no such correlation in the bilingual condition, each speaker having utterances in both languages. It is therefore probable that

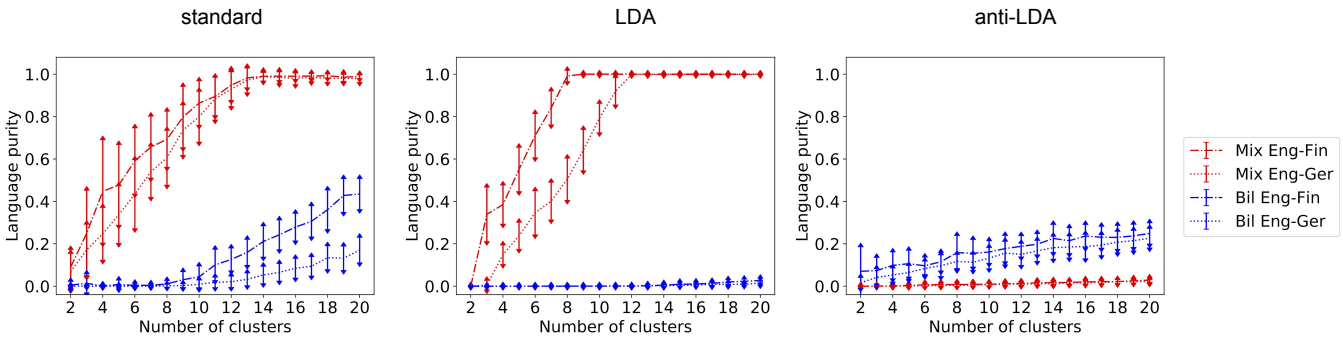

Figure 2: Average language purity as a function of the number of clusters for the different condition, with the standard, LDA and anti-LDA i-vectors. Clustering was done over 20 trials using k-means clustering.

Table 3: Average language purity (in %) using $K = 2$ to $K = 20$ clusters in the different training conditions, with the standard (raw) i-vectors, LDA i-vectors (+lda) and anti-lda i-vectors (-lda).

	English-Finnish			English-German		
Background	raw	+ lda	- lda	raw	+ lda	- lda
Mixed	77.0	83.2	1.3	71.6	68.3	1.0
Bilingual	14.6	0.6	17.0	4.3	0.3	13.6

the acoustic space is clustered primarily by speakers, explaining the highest language purity scores in the mixed condition. Moreover, when the number of clusters is equal to the number of speakers ($K = 12$), clusters in the mixed conditions start reaching perfect purity, while bilingual condition purity scores only start increasing.

The speaker-based cluster hypothesis seems to be confirmed by the results with the LDA and anti-LDA i-vectors. Enhancing speaker information with the LDA favors the mixed condition at the detriment of the bilingual condition, whereas removing this speaker information by taking the LDA's orthogonal complement prevents any language clusters to be formed in the mixed condition, but allows the i-vectors in the bilingual condition to form clusters with language purity scores > 0 when $K < 12$.

It is also worth noting that, in all conditions, the clusters in the English-Finnish dataset have higher purity scores than those in the English-German dataset, suggesting that the language information present in the distant language pair's acoustic space is more discriminatory than those in the close language pair.

We calculated the speaker purity scores for $K = 12$ (the total number of speaker per set). As expected, anti-LDA i-vectors do not cluster speakers at all (M = .011, SD = .003), whereas the LDA i-vectors reach a nearly perfect speaker purity (M = .999, SD = .001). Raw i-vectors also yield very high speaker purity scores (M = .940, SD = .019), suggesting that the standard i-vectors already hold a lot of speaker-specific information.

Discussion

Our experiments successfully replicate the major key findings from previous language discrimination studies, with our model being able to discriminate between languages even with very small exposure. We also found that close language pairs were harder to discriminate than distant ones. However, unlike Carbajal et al. (2016), we found no difference in the standard system between the monolingual and mixed conditions. These results, however, corroborate experimental findings on bilingual children (Byers-Heinlein et al., 2010; Bosch & Sebastián-Gallés, 1997). Although more careful investigation would be required, it is strongly possible that the model in the original study primarily captured recording-specific differences rather than language-specific ones, as the two languages come from distinct datasets. There was also no difference in our raw system with the additional bilingual condition. This would suggest that being exposed to a multilingual environment in which each speaker speaks multiple languages does not hinder the language discrimination process compared to an OPOL-like environment. Although this does not necessarily extend to further processes of language acquisition, such results emphasize the importance of quantitative evidence in supporting psycholinguistics claims.

We found that manipulating the significance of speaker information led to small modulations in language discrimination. Enhanced speaker information slightly improved discrimination in the mixed condition, sometimes to the detriment of the other conditions. Removing this information on the other hand only led to a common very small decrease in discrimination. Such speaker information manipulations assume, in terms of cognition, that infants are able to infer the identity of the speakers in their environments from external modalities (e.g. visual cues). External cognitive processes would then either automatically diminish or enhance speaker-related information when processing speech.

Both theories are also as equally plausible as the standard system, and experimental findings support both ideas: infants are able to recognize speech from their mother (Mehler, Bertoncini, Barriere, & Jassik-Gerschenfeld, 1978) but also fail at strangers voice discrimination tasks when prosody is disturbed (Johnson, Westrek, Nazzi, & Cutler, 2011). Because all three models (raw and with speaker modulation) are reasonable, it would be imprudent to conclude that there are any differences between any of the three exposure conditions.

Findings that bilingual infants are able to discriminate languages (Byers-Heinlein et al., 2010; Genesee, 1989) are not sufficient evidence to assume that they necessarily cluster speech from their multilingual environments into distinct languages. For both mixed and bilingual conditions, and even when manipulating speaker information, the i-vectors used to represent the acoustic space never clustered into two language clusters. This suggests that, even when the number of languages is known, sorting utterances in homogeneous languages clusters is extremely hard. If the number of clusters is increased to the number of speakers, language purity scores increase but only in the mixed condition, corresponding to the intuition of some parents to adopt the OPOL strategy. This indicates that speaker information is not only more salient than the language one, but also that both are intertwined in a way which makes it hard to get them disentangled, even by amplifying or decreasing this speaker information. Nevertheless, it does not mean that these results should be taken as an argument for the OPOL strategy, as there is still no evidence that language separation is a necessary prior to later steps of language acquisition for bilingual children. Hence the underlying question: are children really able to do language diarization? If not, what consequences can it have on language acquisition in bilingual environments?

Another point worth considering in future research is the question of accented speech in bilingual environments. As mentioned previously, the dataset used in the present experiments is composed of non-native bilingual speakers, sometimes leading to the presence of slightly accented English speech. This does not discredit the cognitive inferences made from our results in that even in a family where both parents are native of the two languages, they will often still display accented speech in one language (Major, 1992). However, it would be interesting to replicate the experiments with a corpus solely composed of recordings from native bilinguals, not only to confirm the present results but also to get more insights on the effect of different input types and degrees of accented speech on language discrimination and language diarization.

Acknowledgments

This work was funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute), the CIFAR LMB program, and a grant from Facebook AI Research (Research Grant).

References

Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*(1), 33–69.

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

Byers-Heinlein, K., & Lew-Williams, C. (2013). Bilingualism in the early years: What the science says. *LEARNing landscapes*, *7*(1), 95.

Carbajal, M. J. (2018). *Separation and acquisition of two languages in early childhood : A multidisciplinary approach*. Doctoral dissertation, Université de recherche Paris Sciences et Lettres.

Carbajal, M. J., Dawud, A., Thiollière, R., & Dupoux, E. (2016). The "language filter" hypothesis: A feasibility study of language separation in infancy using unsupervised clustering of i-vectors. In *2016 joint ieee international conference on development and learning and epigenetic robotics (icdl-epirob)* (pp. 195–201).

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association.*

Genesee, F. (1989). Early bilingual development: One language or two? *Journal of child language*, *16*(1), 161–179.

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.

Major, R. C. (1992). Losing english as a first language. *The Modern Language Journal*, *76*(2), 190–208.

Martinez, D., Burget, L., Ferrer, L., & Scheffer, N. (2012). ivector-based prosodic system for language identification. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4861–4864).

Martinez, D., Lleida, E., Ortega, A., & Miguel, A. (2013). Prosodic features and formant modeling for an ivector-based language recognition system. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 6847–6851).

Martinez, D., Plchot, O., Burget, L., Glembek, O., & Matějka, P. (2011). Language recognition in ivectors space. In *Twelfth annual conference of the international speech communication association.*

Mehler, J., Bertoncini, J., Barriere, M., & Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception*, *7*(5), 491–497.

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.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, *116*, 374–388.

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), 1–19.

Oller, D. K., Eilers, R. E., Urbano, R., & Cobo-Lewis, A. B. (1997). Development of precursors to speech in infants exposed to two languages. *Journal of child language*, *24*(2), 407–425.

Petitto, L. A., & Holowka, S. (2002). Evaluating attributions of delay and confusion in young bilinguals: Special insights from infants acquiring a signed and a spoken language. *Sign Language Studies*, *3*(1), 4–33.

Petitto, L. A., Katerelos, M., Levy, B. G., Gauna, K., Tétreault, K., & Ferraro, V. (2001). Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition. *Journal of child language*, *28*(2), 453–496.

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.

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..

Vihman, M. M., Thierry, G., Lum, J., Keren-Portnoy, T., & Martin, P. (2007). Onset of word form recognition in english, welsh, and english–welsh bilingual infants. *Applied Psycholinguistics*, *28*(3), 475–493.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... Contributors, S. . . (2019, Jul). SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, arXiv:1907.10121.

Wester, M. (2010). *The emime bilingual database* (Tech. Rep.). The University of Edinburgh.