Is the Language Familiarity Effect gradual? A computational modelling approach

Maureen de Seyssel, Guillaume Wisniewski, Emmanuel Dupoux

Is the Language Familiarity Effect gradual ? A computational modelling approach

Maureen de Seyssel[12] **(maureen.deseyssel@gmail.com)**
Guillaume Wisniewski[2] **(guillaume.wisniewski@u-paris.fr)**
Emmanuel Dupoux[1] **(emmanuel.dupoux@gmail.com)**
[1] Cognitive Machine Learning (ENS–CNRS–EHESS–INRIA–PSL Research University)
[2] Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle
Paris, France

Abstract

According to the Language Familiarity Effect (LFE), people are better at discriminating between speakers of their native language. Although this cognitive effect was largely studied in the literature, experiments have only been conducted on a limited number of language pairs and their results only show the presence of the effect without yielding a gradual measure that may vary across language pairs. In this work, we show that the computational model of LFE introduced by Thorburn, Feldman, and Schatz (2019) can address these two limitations. In a first experiment, we attest to this model's capacity to obtain a gradual measure of the LFE by replicating behavioural findings on native and accented speech. In a second experiment, we evaluate LFE on a large number of language pairs, including many which have never been tested on humans. We show that the effect is replicated across a wide array of languages, providing further evidence of its universality. Building on the gradual measure of LFE, we also show that languages belonging to the same family yield smaller scores, supporting the idea of an effect of language distance on LFE.
Keywords: language familiarity effect ; computational modelling ; i-vectors

Introduction

The Language Familiarity Effect (LFE) is a cognitive effect observed in language processing, according to which people are better at discriminating speakers who speak in their native language, compared to speakers of another unfamiliar language (Goggin, Thompson, Strube, & Simental, 1991; Johnson, Bruggeman, & Cutler, 2018). Two views are commonly proposed to explain the LFE (T. K. Perrachione, 2018). According to the Phonetic Familiarity hypothesis, the lack of familiarity with the foreign language's lower levels of linguistic characteristics (rhythm, phonetics, acoustics) is enough to explain the effect. For proponents of the Linguistic Processing hypothesis, on the other hand, the effect is in great part explained by the lack of understanding (due to knowledge of lexicon and syntax). However, even in this second view, the role of low-level linguistic features is accepted (Bregman & Creel, 2014; T. Perrachione, Dougherty, McLaughlin, & Lember, 2015).

Methodological issues Although numerous experimental studies run in humans (henceforth behavioural studies) found evidence of the effect, the lack of systematicity makes it hard to compare the results directly (Levi, 2019). First, the evaluation tasks used to assess the presence of LFE differ from one study to the next, ranging from identification tasks (voice line-up) to discrimination tasks (AX task). Critically, a same language pair evaluated on different tasks can yield opposite results regarding the presence of LFE (Levi, 2019). Another source of variability comes from the initial testing conditions. Two setups are principally used, the "1 Group 2 Languages" (or 1G2L) and the "2 Groups 1 Language" (or 2G1L). In the first, most common condition, participants are all native speakers of the same language and are evaluated on their ability to discriminate between speakers in both their native language and a second unfamiliar language. In the 2G1L condition, two groups of participants, native speakers of languages A and B, are tested on only the same language A.

One more issue raised from behavioural studies is the restricted number of language pairs tested. Although this effect has been found over multiple language pairs, leading to qualifying the effect of universal (see Levi (2019); T. K. Perrachione (2018) for reviews), it turns out that only a small number of languages was tested. For instance, only a handful of studies test a language pair that does not contain English (Köster, Schiller, et al., 1997; Johnson, Westrek, Nazzi, & Cutler, 2011; Perea et al., 2014). In order to get more robust evidence of the universality of the effect, a wider array of languages must be tested.

LFE as a gradual effect Because of how the LFE has been evaluated behaviourally, it has mainly been presented as either present or absent. Very few attempts have been made at looking at the effect gradually: T. K. Perrachione (2018) computed effect sizes in LFE experiments, but they are hardly comparable due to differences in setup. Having a systematic gradual measure would allow deeper analyses of specific conditions. Hence, we could directly compare different language pairs or different atypical populations on the LFE. Some studies looked into the role of language distance in LFE. For example, (Levi, 2019) showed, in an extensive literature review, that language pairs both from the same and different rhythmic classes could yield an LFE. However, this does not allow a gradual ranking of language pairs. A few studies directly tested multiple languages with the same population, permitting ranked comparisons, assuming that phonologically similar languages yield better performance in speaker identification. However, these studies never tested more than three languages at a time, and conflicting results were found. Köster et al. (1997) and Zarate, Tian, Woods, and Poeppel (2015) results confirmed this assumption (testing Chinese, English and

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Spanish on German adult listeners and English, German and Mandarin on English adult listeners, respectively). However, no difference between phonologically similar (English and Dutch) and dissimilar (English and Mandarin) were found in infants by (Johnson et al., 2011), thus a need for further studies on the question. Hence, there is a need for a way to test and rank in a systematic manner a large number of language pairs, varying in language similarity.

Additionally, having a gradual measure of the LFE can help analyse finer granularity than that of language differences. An existing example is the case of accented speech. Indeed, some studies found that, for a language pair A-B, if the test stimuli in language A are spoken by native speakers of language B, and therefore accented in B, the LFE can be reduced (Goggin et al., 1991), and even totally cancelled (Goldstein, Knight, Bailis, & Conover, 1981). This suggests that the LFE can be modulated by how heavily accented the speech is and, to a further extent, that acoustically similar dialects should give rise to smaller LFE, corroborating the idea that language distance plays a role in this cognitive effect. These results also show that the effect is gradual, emphasising the need for a gradual measure.

I-vectors as a model of LFE Recently, Thorburn et al. (2019) were able to computationally model the LFE using i-vectors (Dehak et al., 2010), an unsupervised algorithm that allows to compute a representation of whole speech utterances. Computational modelling of LFE can help circumvent some of the methodological problems presented earlier, and we believe it can help compute a systematic, gradual and comparable measure of the effect.

I-vectors models, typically used for speaker-identification applications in speech processing, consist in training a Gaussian Mixture Model on speech features of the train sets utterances to define a new representation of the acoustic space. Then, projecting the components of highest variability onto a lower-dimensional space, we can create a new representation of speech (the i-vector) for all utterances from the train set. By extension, we can predict representations for novel utterances based on this operation. Furthermore, because computed at the utterance level rather than at a finer frame level, we capture the acoustic information that is representative of the utterance as a whole, such as speaker or language information. The lack of time-dependencies in the representation means that only low-level features of linguistics (rhythm, some phonology) are captured. Because of that, and the fact that training such models only necessitates a small amount of input data, the approach has mainly been proposed in models of infants' speech perception. Still, we believe i-vectors can equally model some aspects of adult speech perception that do not require access to higher levels of linguistics.

I-vectors were first proposed in the context of speech perception by Carbajal, Dawud, Thiollière, and Dupoux (2016) as a model of language discrimination. Still modelling language discrimination processes, de Seyssel and Dupoux (2020) showed that they also capture speaker information, even without relying on any supervised components usually present in speech processing applications of i-vectors. Because the LFE depends on both language and speaker information, the i-vector model has the necessary attributes to model it, and this is indeed what showed Thorburn et al. (2019). In their paper, the authors focused on the English-Japanese pair. They showed that the scores from a speaker discrimination task carried out on i-vector representations extracted on both languages were significantly better when the i-vectors were extracted using a model trained on the same language than on another unfamiliar language, effectively replicating the effect found in humans.

Contributions

One underlying contribution of this paper is a replication of the computational approach of Thorburn et al. (2019) on new speech stimuli and language pairs. This reinforces the validity of the i-vector approach to model the LFE. Most importantly, and as the main contribution, we inquire about the capacity of the approach to yield a gradual, comparable measure.

In a first experiment, we look into *reproducing the human findings* according to which accented speech minimises the LFE compared to native speech. Precisely, we replicated an experiment from (Wester, 2012) testing LFE on two language pairs that are always accented in one of the languages. This first experiment also allows us to directly *compare* a close language pair to a distant one. We expect three primary outcomes : a replication of the LFE on the native condition; an LFE is smaller or non-existent in the accented condition compared to the native condition; an LFE is smaller in the close language distance pair than in the distant one. Such results would corroborate with the idea of a gradual effect of the LFE, which we could measure using the i-vector approach, allowing for systematic comparisons.

Findings from the first experiment then lead us to generalise the method to *additional language pairs*. In the second experiment, we evaluate the LFE on 36 language pairs, with many that have never been tested in humans. We can then systematically (1) test and (2) compare the LFE on a large number of languages pairs in a way that would be impossible behaviourally. We expect the LFE to replicate on most of these pairs and to find an effect of language distance.

General methods

The methods presented here are common to both experiments. We use as an example a setup in which we want to evaluate the LFE on a language pair A, B. For each language, we have a set of speech utterances, split between a train and a test set. The former is used to train the models, and the latter is the evaluation stimuli used to test the presence of LFE.

Training pipeline

We first extract Mel Frequency Cepstral Coefficients (MFCCs) (Mermelstein, 1976) for all utterances (train and test), with 13 coefficients including energy, along with double

delta coefficients. We also include pitch information through computation of the fundamental frequency, as it is thought to be relevant in language discrimination (Lin & Wang, 2005).

We then train two i-vector models using the MFCCs from the train sets, one on language A (model A) and one on language B (model B), following the approach first proposed in Dehak et al. (2010). The only difference with the original i-vector approach is that we do not carry out a Linear Discriminant Analysis (LDA), originally aiming at maximising the distance between speakers and/or language (Kanagasundaram, Vogt, Dean, Sridharan, & Mason, 2011; Dehak, Torres-Carrasquillo, Reynolds, & Dehak, 2011). Indeed, as in previous studies using i-vectors as models of speech perception (Carbajal et al., 2016; de Seyssel & Dupoux, 2020; Thorburn et al., 2019), we ensure that the pipeline is unsupervised and therefore better suited to cognitive models. Finally, we extract i-vector representations from the two test sets on both models. That is, we extract i-vectors using model A on tests sets A and B, and similarly for model B. This leaves us with four sets of i-vectors: language A trained on A, language A trained on B, language B trained on A and language B trained on B.

The models are trained with 128 (2,048) Gaussians and i-vectors of dimension 150 (400) in Experiment 1 (Experiment 2). The difference in parameters between the two experiments is explained by the larger number of speakers in the training sets of Experiment 2. Feature extraction, models training and i-vectors extraction were conducted using the Kaldi toolkit (Povey et al., 2011).

Evaluation

Following Thorburn et al. (2019), we first use a machine ABX task (Schatz et al., 2013) to evaluate the capacity of a model to discriminate speakers. In this setup, we create triplets of three utterances from the same language: a, b and x, with a and x being pronounced by the same speaker and b by a different speaker. If the Euclidean distance between the representations (i.e. i-vectors) of utterances a and x is larger than the distance between the representation of b and x, we consider that the model did not manage to discriminate between the speakers, and we count an error for this specific triplet. The ABX error score is the error rate estimated over all possible triplets in the test set.

This framework can be extended to evaluate the LFE by comparing the capacity of a model to discriminate between speakers in a 'familiar' condition, in which the representation is learnt and tested on the same language, to its capacity to discriminate between speakers in an 'unfamiliar' condition, in which test utterances are in a different language than the ones used to train the model. More precisely, we define the LFE score as follows: for a language pair (A,B), we compute the ABX error rates for all four conditions (Ts stands for *test* and Tr for *training*): $\mathrm{Ts}(A)_{\mathrm{Tr}(A)}$, $\mathrm{Ts}(A)_{\mathrm{Tr}(B)}$, $\mathrm{Ts}(B)_{\mathrm{Tr}(B)}$ and $Ts(B)_{Tr(A)}$, where $Ts(A)_{Tr(B)}$ corresponds to the evaluation of the ABX error rate on the language A when the representation has been trained on language B. We then average the scores in the 'familiar' condition ($\mathrm{Ts}(A)_{\mathrm{Tr}(A)}$ and $\mathrm{Ts}(B)_{\mathrm{Tr}(B)}$, the test and train sets being matched in language), and the scores in the 'unfamiliar' condition ($\mathrm{Ts}(A)_{\mathrm{Tr}(B)}$ and $\mathrm{Ts}(B)_{\mathrm{Tr}(A)}$ in which train and test languages are different). The LFE score is defined as the relative percentage increase from the 'familiar' to the 'unfamiliar' condition:

$$\mathrm{LFE} = \frac{S_{\mathrm{diff}} - S_{\mathrm{same}}}{S_{\mathrm{same}}} \tag{1}$$

where:

$$S_{\mathrm{same}} = \frac{\mathrm{Ts}(A)_{\mathrm{Tr}(A)} + \mathrm{Ts}(B)_{\mathrm{Tr}(B)}}{2} \tag{2}$$

$$S_{\mathrm{diff}} = \frac{\mathrm{Ts}(A)_{\mathrm{Tr}(B)} + \mathrm{Ts}(B)_{\mathrm{Tr}(A)}}{2} \tag{3}$$

We use a Two-Sample Fisher-Pitman Permutation Test with Monte-Carlo sampling to test whether this effect is significant. The score is significant if discrimination scores in the S_{same} and S_{same} groups are significantly different. A positive significant LFE score reflects an effect of language familiarity, with a higher ABX error rate in the non-familiar condition than in the familiar condition.

Because we are looking at the LFE on the language pair symmetrically, that is analogous to a '2 groups 2 languages' (or '2G2L') approach, (two groups of participants, native in two different languages, are tested on both languages). Hence, we are controlling for any biases due to a specific training set yielding better speaker discrimination performance, and thus singling out the actual LFE process. This is a more robust evaluation setup than what is commonly done in behavioural work, where LFE is looked into from the perspective of a single language only.

Experiment 1: LFE and accented speech

First, we focus on two language pairs, English-Finnish and English-German. For each of these pairs, we compare a "native" setup, where all tested speakers are native in the languages, and an "accented" setup, with English utterances being spoken by non-native speakers, hence Finnish accented or German-accented.

Materials

We retrieved audiobooks in English, German and Finnish from the LibriVox project[1] using the `Libri-Light` tools (Kahn et al., 2020), and used a Voice Activity Detection model (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020) to segment speech. We then created for each language a 10 hours training set, balanced equally between 10 speakers.

The test sets were built from the EMIME bilingual corpus (Wester, 2010), which contains English, German and Finnish read speech uttered by native speakers, as well as English spoken by German and Finnish speakers, and is therefore accented. We built five different test sets: native Finnish, native German, native English, Finnish-accented English and

[1] https://librivox.org

Table 1: Summary of test sets in Experiment 1.

Language	Accent type	N speakers (N male)	Mean (SD) utt dur (in s)
English	native	12 (6)	3.21 *(1.04)*
	Finnish	12 (6)	4.37 *(1.48)*
	German	12 (6)	4.56 *(1.52)*
Finnish	native	12 (6)	4.6 *(1.29)*
German	native	12 (6)	4.6 *(1.32)*

German-accented English. Each test set is balanced equally between 12 speakers and has an average duration of 25 min (348 utterances). See Table 1 for more information.

Results

We calculated the LFE score on four language pairs following the procedure presented in the General Methods: native English and native German; native English and native Finnish; German-accented English and native German; Finnish-accented English and native Finnish. We refer to the two first pairs as *native* and the two last as *accented*.

Table 2: LFE scores on the native and accented conditions for both language pairs in Experiment 1. Significance was estimated using a Two tailed Paired Fisher-Pitman Permutation Test with Monte-Carlo sampling (*: p<.05)

Language Pair	LFE (%)	
	native	accented
English - Finnish	**+19.21***	-8.62
English - German	**+10.77 ***	-1.1

The first thing to notice from Table 2 is that both language pairs yield a *significant* LFE score in the native condition (the familiar models yield better discrimination scores than the unfamiliar ones, p<.05 in both pairs), giving further support to the i-vector approach as a good model of LFE. Moreover, the LFE score is higher in the English-Finnish pair than in the English-German pair, suggesting that the distance between languages could modulate the LFE.

In the accented condition, there is no longer a significant difference between the familiar and unfamiliar models' scores on language discrimination, and this on both language pairs. Hence, whilst the LFE scores indicate the effect was present in the native conditions, it is no longer the case in the accented condition, that is, when one of the two languages is uttered with an accent from the other language. These results, which replicate the behavioural findings from Wester (2012) as well as previous studies on accents, suggest that we can use the i-vector models to obtain a gradual measure of the LFE.

Experiment 2: Testing LFE on many language pairs

Results from the first experiment not only validate further the i-vector approach as a good model of the LFE, but they also suggest that the resulting measure is gradual and thus comparable. Furthermore, they suggest that there might be an effect of language distance on LFE.

In this second experiment, we generalise the experiment to many language pairs, including pairs that have not been tested on humans. This allows us to 1) verify the universality of the effect, 2) make use of the gradual measure to compare pairs with varying language distances.

Materials

We used stimuli from the `CommonVoice 6.1` (CV) corpus (Ardila et al., 2019), which gathers read speech from a large number of languages. We selected nine languages (those for which we had enough data) and generated, for each of them, training sets of 15 hours split between 60 speakers and test sets of 30 minutes split between 20 speakers (see Table 3 for the complete list). The high number of speakers is closer to the setup proposed by Thorburn et al. (2019) than in the first experiment and ensures more variability in the training set, leading to a more robust model.

Table 3: Summary of languages in Experiment 2. Train sets have an average duration of 15 hours (60 speakers) and test sets have an average total duration of 30mn (20 speakers).

Language	**ISO**	**Avg utt dur (s)**	**Family**
Catalan	cat	5.10 *(SD=1.82)*	indo-european
Welsh	cy	4.52 *(SD=1.67)*	indo-european
German	deu	4.42 *(SD=1.50)*	indo-european
English	eng	4.81 *(SD=1.74)*	indo-european
Farsi	fas	3.80 *(SD=1.42)*	indo-european
French	fra	4.78 *(SD=1.51)*	indo-european
Italian	ita	5.35 *(SD=1.73)*	indo-european
Kabyle	kab	3.38 *(SD=1.23)*	afro-asiatic
Kinyarwanda	kin	5.14*(SD=1.80)*	niger-congo

Results

Models were trained on the nine languages, and evaluation was run on all possible language pairs, yielding 36 LFE scores.

Speaker ABX scores averaged across all pairs are presented in Figure 1, and detailed LFE scores are available in Table 4. Speaker discrimination scores are overall significantly higher in the 'familiar' condition than in the 'unfamiliar' one, with a mean LFE of 13.78 (significance was calculated using a 95% confidence interval with bootstrapping on languages, with 10,000 permutations, CI = [2.71–18.55]). These results corroborate the idea of a universal LFE that can be expected on language pairs that were not tested on humans.

However, the difference is not systematically significant in every pair, with one language pair (German-Welsh) yielding a significant inverse LFE.

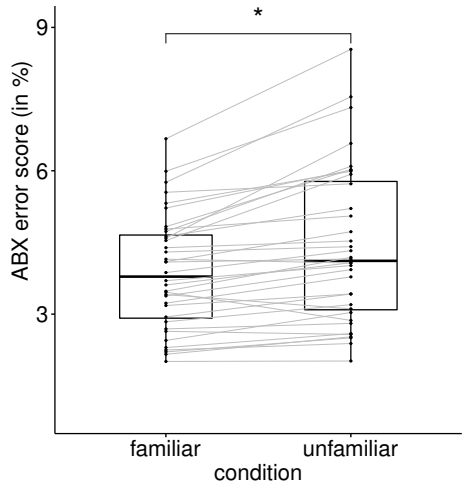

Figure 1: Speaker ABX error scores averaged across the 36 language pairs in the CV dataset. LFE score = 13.78. The asterisks on top illustrate the significance level (*, 95% CI).

We then divided the language pairs into two groups: the 'same family' and the 'different family', based on whether the languages in the pair belong or not to the same language family (as defined by the WALS typology (Dryer & Haspelmath, 2013), see Table 3). As shown in Figure 2, the LFE scores from 'same family' pairs (M=21.46, SD=9.62), N=15 are significantly lower than the 'different family' pairs (M=6.13, SD=9.47, N=21) (significance was tested using a 99% confidence interval with bootstrapping on language within family with 10,000 permutations, CI = [7.28,29.07]).

Discussion

In the first experiment, we successfully replicated results from Thorburn et al. (2019) by showing that the i-vector approach yields a significantly positive LFE score on two new language pairs (native condition). Moreover, we further validated this model by replicating another behaviour observed in humans, that is, the fact that the LFE can be diminished or cancelled with accented speech (Goldstein et al., 1981; Thompson, 1987). Specifically, the stimuli we use in our 'accented' conditions come from the same dataset as in (Wester, 2012), in which they also found no significant effect of LFE in humans, neither in the English-German nor in the English-Finnish pair.

Differences in the LFE between accented and native speech support the idea of the LFE as a gradual effect that can be modulated by languages variations. Current psycholinguistic setups make such changes hard to capture in humans, but our results suggest that with the i-vector approach, we can capture and grade such changes. Being able to capture this granularity allows one to investigate the role of different variables

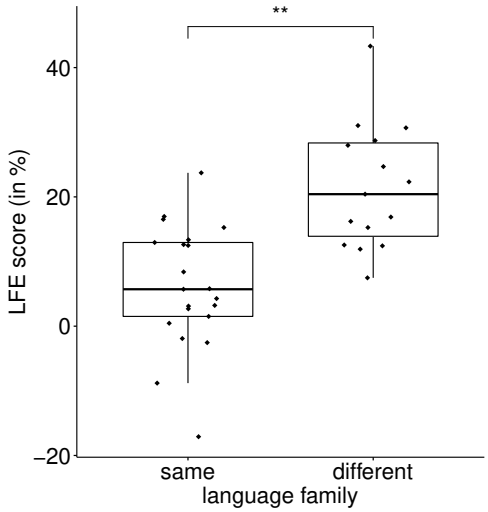

Figure 2: LFE scores averaged across the 'same family' and 'different family' conditions. The asterisks on top illustrate the significance level (**, 99% CI).

on the LFE, with the most obvious one being that of language distance. As discussed earlier, the role of language distance is hard to analyse in behavioural experiments. However, results from Experiment 1 do suggest an effect, with the close distance language pair (English-German) yielding a lower LFE than the distant language pair (English-Finnish).

In the second experiment, we tested the model on a larger number of language pairs, of which many have never been tested on humans. We could compute comparable LFE scores for each of the 36 language pairs. We note that the LFE was present overall (across all pairs together), validating the approach again. However, not all language pairs yielded a significant effect. Multiple things could cause this: these specific pairs might not actually yield an LFE in humans, or the LFE might be too small in humans, and the model is not sensitive enough to capture it. Regardless, we should replicate the experiment behaviourally, especially if the pair had not been tested on humans before. Finally, there might also be specific biases in the stimuli resulting in an absence of LFE. For example, two pairs, Welsh-English and Welsh-German, actually yielded a negative LFE score: the unfamiliar condition yielded better discrimination scores than the familiar one. However, it is likely that a large part of the Welsh utterances in the CV corpus was pronounced by English native speakers, which, in light of the previous results on accented speech and LFE, could partially explain the lack of LFE.

Interesting results arose when the language pairs were divided into two groups: those in which both languages of the pair belong to the same family and those in which they do not. There was a significant difference in LFE scores between the two groups, with the same family language pairs yielding much lower LFE scores than the different family pairs. This corroborates with results from Experiment 1, suggesting that

Table 4: LFE scores for all possible CommonVoice language pairs. Two tailed Paired Fisher-Pitman Permutation Test with Monte-Carlo sampling with Bonferroni correction (*: p<.05; **: p<.005)

	ca	**cy**	**de**	**en**	**fa**	**fr**	**it**	**kab**
cy	8.39							
de	12.49	*-17.10***						
en	5.80	-8.81	0.45					
fa	**16.53***	1.51	4.27	-2.55				
fr	**23.72****	12.95	**13.36****	**16.97***	**12.63****			
it	3.11	3.21	-1.91	**15.27****	5.71	2.68		
kab	**20.42****	12.42	**16.88****	7.48	**11.92****	**16.21****	**12.55****	
rw	**24.69****	**22.32****	**43.32****	**30.68****	**15.26****	**28.71****	**27.97****	**31.04****

closer languages lead to a smaller LFE. The possibility for the LFE to be affected by language distance raises many interesting points regarding the cognitive processes behind this phenomenon. Commonalities and differences between languages can occur at various linguistic levels, but the i-vector approach only focuses on low-level cues (mainly phonology, prosody and phonotactics). Yet, it suggests an effect of language distance, supporting further the idea that the LFE is largely due to the familiarity with the phonetics and phonology of the foreign language, as proposed by the *phonetic familiarity* hypothesis (Fleming, Giordano, Caldara, & Belin, 2014; Orena, Theodore, & Polka, 2015). Still, the distance between two languages at higher linguistic levels may also enhance the phenomenon.

To conclude, although it is not guaranteed that the language distance effect suggested by the model is equally present in humans, our results give us strong incentives to investigate this. This should be done in a systematic setup allowing for direct comparison of language pairs, potentially by designing a wide-scale online speaker discrimination study in many languages. Still, it is yet unclear whether we can obtain a fine enough gradual measure in humans.

LFE score stability One of the central issues in computational modelling is the impact of data on the models. Here, we consider that the i-vector-based LFE score is stable in that it is not affected by changes in train or test sets. This is why we can confidently compare two conditions (languages, setup, number of speakers, recording condition) as long as all other factors are controlled for. However, we have not tested whether the results are prone to variations based on the train and evaluation stimuli, for example by running the same experiments on a new train or test sampled from the same original dataset. Only if the results are stable can we fully validate the approach. This stability aspect also raises the question of how representative of a language the training sets are. Indeed, while humans have had years of being exposed to their native language, which allows them to build an internal language prototype, we only train the models on a few hours of data in the current approach. Despite being a considerable advantage in data collection, it also increases the probability for the model's prototype to be biased. In the second experiment, we purposely used a high number of speakers and diversity in the recording setup, and we recommend any further work to follow this lead.

Finally, we would like to address the fact that the i-vector model was initially proposed as a model of infant perception (Carbajal et al., 2016; de Seyssel & Dupoux, 2020), and used as such in the scope of the LFE to support the evidence that the effect only requires low-level linguistic knowledge and is present in infants (Thorburn et al., 2019). Here, we focused on the LFE in general, without restriction to a specific age group. Indeed, although the present model only requires knowledge of the acoustics of the language, it still reproduces behavioural results found in adults, and we can only assume that adding higher up knowledge that makes use of language's understanding will only reinforce the effect found here. Therefore, even if the model could be completed by adding such features, the present approach can still be seen as a model of LFE in both infants and adults and has the advantage of only requiring very little data.

Conclusion

To conclude, our results further validate the i-vector approach as a good model of the LFE by replicating Thorburn et al. (2019) on novel languages and replicating human experiments on accented speech. These results on accents also suggest that the effect can be modulated, hence gradual. The i-vector model allows computation of gradual LFE scores, meaning that we can then directly compare different conditions. We showed further evidence of the universality of the effect by evaluating it on a large number of pairs systematically, in a way that can only be done computationally. We also found an effect of language distance, with larger LFE yielded when the two languages are dissimilar. These results should be replicated with humans in a setup allowing systematic evaluation, and attention should be given to the design of such an experiment. Finally, a more thorough analysis of the stability of the model depending on the training set could be done to ensure that the scores are fully stable and that different data would not give different scores, skewing the comparisons.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, *130*(1), 85–95.

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., Dehak, R., Glass, J. R., Reynolds, D. A., Kenny, P., et al. (2010). Cosine similarity scoring without score normalization techniques. In *Odyssey* (p. 15).

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.*

de Seyssel, M., & Dupoux, E. (2020). Does bilingual input hurt? a simulation of language discrimination and clustering using i-vectors. In *Cogsci 2020-42nd annual virtual meeting of the cognitive science society.*

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/`

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014, September). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, *111*(38), 13795–13798. Retrieved 2021-01-29, from `https://www.pnas.org/content/111/38/13795` (Publisher: National Academy of Sciences Section: Social Sciences) doi: 10.1073/pnas.1401383111

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.

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society*, *17*(5), 217–220.

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.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., ... others (2020). Libri-light: A benchmark for asr with limited or no supervision. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7669–7673).

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., & Mason, M. (2011). I-vector based speaker recognition on short utterances. In *Proceedings of the 12th annual conference of the international speech communication association* (pp. 2341–2344).

Köster, O., Schiller, N. O., et al. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics*, *4*, 18–28.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*.

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(2), e1483.

Lin, C.-Y., & Wang, H.-C. (2005). Language identification using pitch contour information. In *Proceedings.(icassp'05). ieee international conference on acoustics, speech, and signal processing, 2005.* (Vol. 1, pp. I–601).

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

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition*, *143*, 36–40.

Perea, M., Jiménez, M., Suárez-Coalla, P., Fernández, N., Viña, C., & Cuetos, F. (2014). Ability for voice recognition is a marker for dyslexia in children. *Experimental Psychology*.

Perrachione, T., Dougherty, S., McLaughlin, D., & Lember, R. (2015). The effects of speech perception and speech comprehension on talker identification. In *Icphs.*

Perrachione, T. K. (2018, December). Recognizing Speakers Across Languages. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 514–538). Oxford University Press. doi: 10.1093/oxfordhb/9780198743187.013.23

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others (2011). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.*

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..

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121–131.

Thorburn, C. A., Feldman, N. H., & Schatz, T. (2019). A quantitative model of the language familiarity effect in infancy. In *Proceedings of the conference on cognitive computational neuroscience.*

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

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781–790.

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific reports*, *5*(1), 1–9.