INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

The ART of Conversation: Measuring Phonetic Convergence and Deliberate Imitation in L2-Speech with a Siamese RNN

Zheng Yuan[1,2], *Aldo Pastore*[1,2], *Dorina de Jong*[1,2], *Hao Xu*[3], *Luciano Fadiga*[1,2], *Alessandro D'Ausilio*[1,2]

[1]Istituto Italiano di Tecnologia, Italy
[2]Università degli Studi di Ferrara, Italy
[3] University of California San Diego, USA

zheng.yuan@iit.it

Abstract

Phonetic convergence describes the automatic and unconscious speech adaptation of two interlocutors in a conversation. This paper proposes a Siamese recurrent neural network (RNN) architecture to measure the convergence of the holistic spectral characteristics of speech sounds in an L2-L2 interaction. We extend an alternating reading task (the ART) dataset by adding 20 native Slovak L2 English speakers. We train and test the Siamese RNN model to measure phonetic convergence of L2 English speech from three different native language groups: Italian (9 dyads), French (10 dyads) and Slovak (10 dyads). Our results indicate that the Siamese RNN model effectively captures the dynamics of phonetic convergence and the speaker's imitation ability. Moreover, this text-independent model is scalable and capable of handling L1-induced speaker variability.

Index Terms: phonetic convergence, Siamese RNN, speech imitation, alternating reading task, L2 English

1. Introduction

Phonetic convergence is a phenomenon in which speakers modify their acoustic-phonetic repertoire to approximate that of their conversational partners [1] for various purposes such as enhancing social affiliation, facilitating communication [2, 3], expressing identity or improving language proficiency [4], sometimes unconsciously or without a clear intention [5].

Phonetic convergence also reflects the complex interactions between social and cognitive factors that influence speech perception and production [3, 6, 7]. It can reveal how speakers store and process linguistic representations in the brain, how they monitor their own and others' speech, and how they use social cues to guide their linguistic behaviour [8, 9]. It also sheds light on the mechanisms of sound change and language evolution [10].

In the context of language acquisition, phonetic convergence can affect L2 development positively by allowing learners to adjust to native-like pronunciation and prosody or negatively by strengthening their L1 features or non-native features adopted from other L2 speakers [11, 12, 13]. Despite extensive research on phonetic convergence involving native speakers [10, 14, 15], the question of whether such convergence also occurs between non-native speakers remains largely unanswered [16]. To study the dynamics of phonetic convergence in L2-L2 interaction, we designed the Alternating Reading Task (ART)[17], a scripted text reading experiment, based on previous studies[18, 19, 20, 21, 22, 23]. ART maintains the turn-taking structure in natural conversations and provides a controllable experimental complexity. According to past research, a speaker's ability to imitate can be an important factor for phonetic convergence (implicit imitation) [6, 9, 24], we incorporated an explicit imitation condition in our experiment (see Section 2). The ART dataset that we collected comprised L2 English speech data from native speakers of Italian, French, and Slovak.

How to measure the degree of phonetic convergence remains an open area of research, with work featuring both subjective evaluations [1, 6, 25] and objective modelling [20, 23, 26]. In this work, we focus on modelling the holistic convergence as it is a more direct indicator of the interaction between speech perception and production [25]. It is measured by globally comparing temporal and spectral characteristics of two speech signals[27]. The classic speaker verification method Gaussian Mixture Model-Universal Background Model (GMM-UBM) [28] proved robust in assessing holistic phonetic convergence at the word[19, 20, 21, 22] and sentence level [17] by comparing the log-likelihood ratio (LLR) of a speaker's baseline and interactive condition. However, GMM-UBM has its limitations when the speaker number grows large with increasing dialect, language and speaker variability. In this paper, we present a Siamese neural network [29] with a recurrent architecture (Siamese RNN) [30], to measure phonetic convergence and speech similarity [1], exploiting the neural network's capacity to learn intricate and non-linear speaker feature representations [31]. To our best knowledge, Siamese neural networks have not been adopted and experimented on phonetic convergence studies.

2. Alternating Reading Task dataset

2.1. Participants

We recruited 58 participants for the ART experiment: 20 native French (all female, average age 23.45±4.94), 18 native Italian (6 males, average age 24.50±3.65) in the initial version[32], and 20 native Slovak (10 males, average age 33.75±13.69) for this phase[33]. To ensure B2-level English reading competency, all participants passed an online proficiency test [34] (test score: French=82.59±9.53, Italian=74.16±6.70, Slovak=78.12±10.24). Same-sex dyads were formed with participants having similar reading proficiency ($<15\%$ difference in test scores), and most of them did not know each other prior to the experiment.

2.2. Task description

In the ART experiment[17], dyads took turns reading aloud a neutral English text. The text contained 80 sentences and turn boundaries were set within sentences. We replaced some words with synonyms to maintain attention. To make comparisons, a solo and an imitation condition were added. In the solo session, participants read the sentences individually, and in the interactive condition, participants performed the Alternating Reading

[1]The code is available at https://github.com/byronthecoder/S-RNN-4-ART.

10.21437/Interspeech.2023-2283

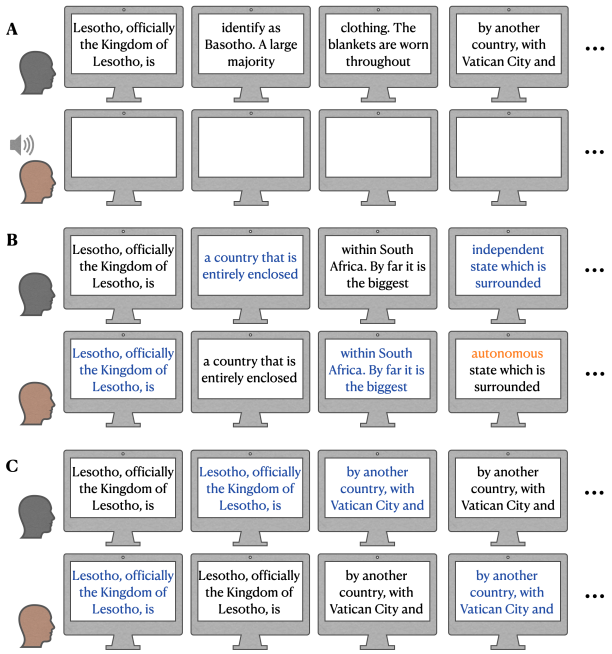

Figure 1: *The Alternating Reading Task. Participants speak when their computer screen shows black sentences. (A) The solo condition; (B) the interactive condition - synonym shown in orange for illustrative purpose only, and (C) the imitation condition.*

Task four times with varying degrees of word replacement. The final imitation session required participants to imitate each other during sentence reading. Participants could not see each other during the experiment (see Figure 1).

3. Siamese RNN model

We implemented a Siamese recurrent neural network based on Mueller and Thyagarajan's paper [30] to predict speech utterance similarity and capture speaker information via a binary speaker verification task (see Section 4). This algorithm learns a vector representation for each utterance that encodes the acoustic characteristics of the speaker, allowing for direct measurement of phonetic convergence between interlocutors. Additionally, the RNN's ability to preserve historical information during computation leads to refined modelling of speech effects like coarticulation.

As shown in Figure 2 , we employed a two-network structure, with each RNN network processing one audio input. The structure utilizes tied weights, meaning that the trainable parameters in RNN_1 and RNN_2 are shared. Each input speech sentence $s \in S$ consists of a set of fixed length vector $\{\mathbf{x}_t^s\}_T$, where S denotes the corpus, and time step $t \in \{1, ..., T\}$, obtained from MFCC extraction (see 4.1). The network maps the input vector of dimension d_{in} to a representation space of dimension d_{rep} ($d_{in} = 39$ and $d_{rep} = 50$ in our experiments).

Going into the bi-directional RNN layer, the input vector $\mathbf{x}_t$ at time t is used to compute both the forward and backwards hidden states $\overrightarrow{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$. The computation for each sentence $s \in S$ is as follows:

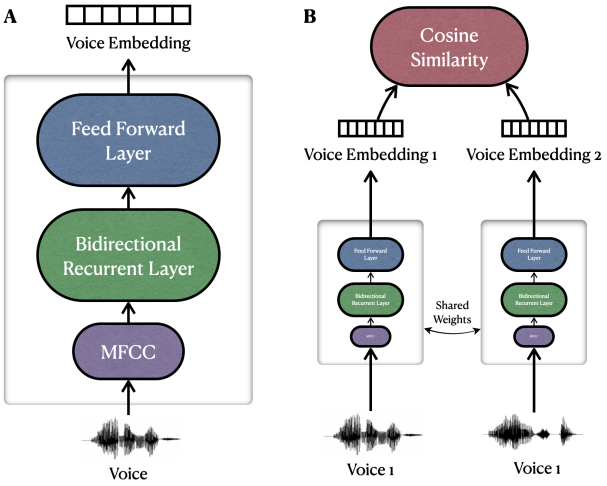

Figure 2: *(A) The proposed voice representation module consists of an MFCC extraction module (purple), a bi-directional RNN layer (green), and a feed-forward layer (blue). It generates a low-dimensional vector representation for input voice audio. (B) shows our model pipeline. The model inputs two voice audios and computes the distance between their voice embeddings from two tied-weight voice representation modules.*

$$\overrightarrow{\mathbf{h}}_t^s = \tanh(\overrightarrow{\mathbf{W}}_h \mathbf{x}_t^s + \overrightarrow{\mathbf{U}}_h \overrightarrow{\mathbf{h}}_{t-1}^s + \overrightarrow{\mathbf{b}}_h), \quad (1)$$

$$\overleftarrow{\mathbf{h}}_t^s = \tanh(\overleftarrow{\mathbf{W}}_h \mathbf{x}_t^s + \overleftarrow{\mathbf{U}}_h \overleftarrow{\mathbf{h}}_{t+1}^s + \overleftarrow{\mathbf{b}}_h), \quad (2)$$

where $\mathbf{W}_h$ and $\mathbf{U}_h$ are weight matrices for all speech segment vectors $\{\mathbf{x}_t\}_T$ and hidden-state vectors $\mathbf{h}_t$ respectively, and $\mathbf{b}_h$ denotes the bias. The hidden-state vectors are initiated as zero vectors $\overrightarrow{\mathbf{h}}_1^s = \overleftarrow{\mathbf{h}}_T^s = \mathbf{0}$. Next, we concatenate the forward and backwards hidden states to obtain the hidden state $\mathbf{h}_t^s = [\overrightarrow{\mathbf{h}}^s, \overleftarrow{\mathbf{h}}^s]$, and fed it into a feedforward layer activated by the $\tanh$ function:

$$\mathbf{y}_t^s = \tanh(\mathbf{W}_y \mathbf{h}_t^s + \mathbf{b}_y), \quad (3)$$

where $\mathbf{W}_y$ and $\mathbf{b}_y$ are the weight matrix and the bias vector associated with the output $\mathbf{y}_t^s$. Additionally, a masking layer has been set before the bi-directional RNNs to cope with sequences of different lengths.

After processing the last time step T, the output on sentence s is $\mathbf{y}_T^s$. Then, the output will go through a feedforward layer activated by the sigmoid function and get transformed into the speech embedding vector $\mathbf{e}^s$ via

$$\mathbf{e}^s = \text{sigmoid}(\mathbf{W}_e \mathbf{y}_T^s + \mathbf{b}_e), \quad (4)$$

where the weights and bias are denoted by $\mathbf{W}_e$ and $\mathbf{b}_e$.

The final step is computing the distances between two speech embedding vectors. In this study, we use cosine similarity to measure the distance. Given the speech embeddings $\mathbf{e}^s$ and $\mathbf{e}^{s'}$ for sentence $s, s' \in S$, we compute

$$g(\mathbf{e}^s, \mathbf{e}^{s'}) = \frac{\mathbf{e}^s \cdot \mathbf{e}^{s'}}{|\mathbf{e}^s|_2 |\mathbf{e}^{s'}|_2} \quad (5)$$

Table 1: *Model performance on different architectures, training data size and test scenarios*

Model	Epoch	Size(K)	Param(K)	Acc
LSTM+FF	10	91.6	**20.7**	0.79
RNN+FF	10	91.6	7.2	0.66
Bi-RNN+FF	10	91.6	14.3	**0.84**

(a)

Model	Epoch	Size(K)	Acc
VIFS10	20	4.2	0.80
VIFS10	100	4.2	**0.87**
VIFS10	150	4.2	**0.87**

(b)

Model	Subset	Acc
VCTK	IT	0.70
VCTK	FR	**0.82**
VCTK	SK	0.75

(c)

Model	Size(K)	Epoch	Acc
IF	31.6	10	0.85
IFS	31.6	10	**0.90**

(d)

4. Speaker verification task

To capture convergence, we trained a Siamese RNN model using a binary speaker verification task. The model predicts whether two speech utterances are produced by the same speaker based on a cosine similarity score. In our ART dataset, the same sentences were produced across the three experimental conditions. We assume that there is a gradual increase in convergence from the solo to the interactive and imitation conditions. A correct prediction that a sentence pair is produced by different speakers, along with an increased similarity score, indicates phonetic convergence. Likewise, speakers in the imitation and interactive conditions should be dissimilar to their own solo baseline, resulting in a decrease in intra-speaker similarity score.

4.1. Data preparation

To build the training and test datasets, we first segmented the audio files into phrase tokens with a Psychopy3 script based on timestamps. Non-voiced segments and noises like laughs and coughs were cut out, but stutters and repeats within the sentences were included. Stereo audios have been converted to mono audios. 39-dimensional MFCC features (13 static with Δ and $\Delta\Delta$) were extracted every $10\,\mathrm{ms}$ with a $25\,\mathrm{ms}$ window size using the Python package Librosa [35]. Cepstral Mean and Variance Normalization were performed to mitigate the mismatch between the recording devices and the variation of the recording environment.

We used the solo data to train, validate and test the model and then further test the model performance on the interactive and imitation data. For each speaker, we created positively labelled data and negatively labelled data. A positive data example was a sentence pair produced by the same speaker, likewise the negative one from different speakers. We only used the real interactive speaker pairs to build the datasets. Surrogate speaker combinations have not been adopted.

The training set was constructed using the first 40 sentences of the script, the validation set using the following 20 sentences, and the test set using the remaining 20 sentences. Using this approach, we generated 45,240 positive data examples and 46,400 negative data examples for the training set. For the validation and test sets, we created 11,020 positive data examples and 11,600 negative data examples each. The data split ratio was thus approximately 70%:15%:15%. The dataset comprises a balanced proportion of participants with diverse L1 backgrounds, and the label distribution is approximately equal across all groups.

To evaluate the model in interactive and imitation conditions, we constructed another test set by extracting a subset of sentence combinations. Positive samples were created using *adjacent sentences* spoken by the same speaker, while negative samples were created using the *same sentence* spoken by different speakers. This arrangement enables phonetic convergence and imitation measurement in Section 5. To build the interactive set, we conducted four rounds of ART experiments. Given the stable phonetic convergence observed in our previous work [17]across four sessions, we randomly selected two sessions with different initial speakers.

4.2. Training Details

Our Siamese RNN model used 50 nodes in both the RNN and feedforward layers. The model trainable parameters included $\mathbf{W}_{h/y/e}$, $\mathbf{U}_h$, and $\mathbf{b}_{h/y/e}$, with a total size of 14,250 and were initialised with small random normal-distributed values. Weight optimization was performed using the Adam [36] method, together with a learning rate decay strategy. Dropout (rate=0.2, after the RNN layer), l_1 regularization and batch-normalization were employed to avoid overfitting and stabilize the training process. The model's loss function was binary cross-entropy and the evaluation metrics were binary accuracy, precision, recall, F1-score and AUC. Our best result was achieved by initializing the model with weights pre-trained on the VCTK corpus [37], a large spoken sentence dataset with 109 L1 English speakers.

4.3. Results

Table 1 summarises our experimentation on different model architectures, training data size and test scenarios which will be discussed in Section 6. In Table 2, we report the precision, recall, F1-score, and AUC of our best-performing model for the binary speaker verification task. The first column shows the ART experimental condition under which the test was conducted. The results indicate that the model performs best in the solo condition, achieving an F1-score of 0.95 for the positive samples and 0.94 for the negative. The model's performance slightly drops in the interactive and imitation conditions as the speech characteristics are unconsciously or deliberately modified in the new scenarios.

Table 2: *Evaluation results of the best model*

	Precision	**Recall**	**F1-score**	**AUC**
Solo				
positive	0.93	0.97	0.95	0.99
negative	0.96	0.92	0.94	0.99
Interactive				
positive	0.81	0.87	0.84	0.91
negative	0.86	0.80	0.82	0.91
Imitation				
positive	0.82	0.86	0.84	0.91
negative	0.85	0.81	0.83	0.91

5. Phonetic convergence measurement

We computed the average similarity scores of all speakers across three conditions with false predictions and outliers removed. Results show that either the intra-dyad similarity score of the imitation condition (0.081±0.092) or that of the interactive condition (0.074±0.085) is comparatively higher than in the solo condition (0.024±0.028). Meanwhile, the intra-speaker cosine similarity shows an opposite tendency with a gradual decrease from the solo condition (0.958±0.052), to the interactive (0.890±0.126) and the imitation condition (0.871±0.139). The results agree with our assumption about the dynamic of phonetic convergence across the experimental conditions. The intra-speaker similarity change is more significant than that of the dyadic setting.

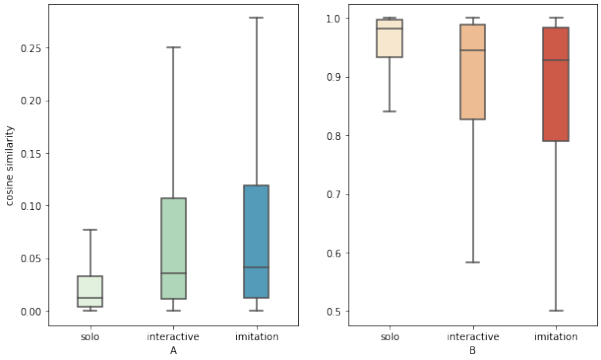

Figure 3: *The cosine similarity scores across the solo, interactive and imitation conditions. Subplot (A) shows intra-dyad similarity scores (different speakers) and (B) the intra-speaker similarity scores (same speaker).*

We also tested whether a better speech imitator has a higher convergence degree in interaction. The imitation ability was defined by the average change of intra-speaker similarity across the solo and imitation conditions. The degree of phonetic convergence was measured as a speaker's average similarity change across the solo and interactive conditions. Figure 4 suggests a positive correlation between the imitation ability and the convergence degree (Pearson correlation coefficient r=0.51, p=0.0005) in the interactive condition. The convergence measurement result shown in Figure 3 and the correlation found in Figure 4 are consistent with our previous work[17] using a GMM-UBM model in similar tasks.

6. Discussions

A simple architecture with fine scalability. The Siamese RNN model has shown its efficacy in speaker verification and convergence measurement tasks, yielding consistent results compared to the traditional GMM-UBM model, despite the notable model differences. The Siamese RNN model is characterized by its light-weighted structure and small parameter size, which facilitate its training and adaptation. We investigated more complex architectures such as LSTMs without pre-training, but they did not outperform the bi-directional RNN (Table1a). Additionally, we evaluated the model's scalability with less training data. We built a subset using 10 sentences and retrained the model based on the pre-trained VCTK weights. The binary accuracy in 20 epochs reached 0.80 and increased to 0.87 in 100 epochs, as shown in Table 1b. Furthermore, we increased the number of speakers in the dataset and found that the model's performance

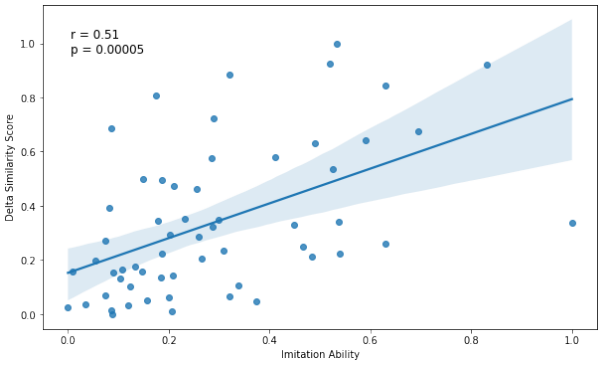

Figure 4: *The correlation between imitation ability and the degree of convergence in interaction with a 95% confidence interval. The blue dots represent speakers. The imitation ability score on the x-axis and the cosine similarity change during interaction on the y-axis were normalized to (0, 1).*

did not decrease as discussed in the following paragraph.

Adaptation to large speaker space to tackle L1-induced variability. In our study, we observed a performance gap when testing the pre-trained VCTK model on L1 Italian, French, and Slovak data, indicating L1-induced data variability(Table 1c). To experiment on the model's ability to handle such variability, we randomly selected 5 female dyads from the Italian and French data pools to build training and validation sets. We trained two models, one with 5 additional random Italian and French female dyads and the other with 5 Slovak female dyads, and repeated the experiment five times to compare the models' average binary accuracy. Table 1d reveals that the model successfully captured the general acoustic features of the speakers. The introduction of new Slovak data did not confuse the model and its performance even improved.

Subjective evaluation and feature engineering as next steps for model improvement. The implementation of the Siamese architecture in this work has limitations as the interpretability of the phonetic convergence score is arguable due to the lack of ground truth. Therefore, subjective evaluation, such as AXB test[1], is necessary for future study of the model. Additionally, a good result was achieved with a relatively agnostic set of acoustic features such as MFCCs, the model's performance in the interactive and imitation condition could be improved by more refined feature engineering or the introduction of linguistic knowledge.

7. Conclusions

This paper presents a novel approach to studying phonetic convergence and speech imitation in L2-L2 interaction, utilizing a Siamese RNN model. The model was validated through a structured and controllable experiment, the scripted text alternated reading task (ART). In addition, the paper expanded the sample diversity by including Slovak L2 English speakers, which investigated the model's scalability and capability to handle L1-induced speaker variability.

8. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 859588.

9. References

[1] J. S. Pardo, "On phonetic convergence during conversational interaction," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[2] H. Giles, N. Coupland, and J. Coupland, "Accommodation theory: Communication, context, and consequence." 1991.

[3] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[4] N. Lewandowski and M. Jilka, "Phonetic convergence, language talent, personality and attention," *Frontiers in Communication*, vol. 4, p. 18, 2019.

[5] J. S. Pardo, "Measuring phonetic convergence in speech production," *Frontiers in psychology*, vol. 4, p. 559, 2013.

[6] S. D. Goldinger, "Echoes of echoes? an episodic theory of lexical access." *Psychological review*, vol. 105, no. 2, p. 251, 1998.

[7] A. M. Liberman and D. H. Whalen, "On the relation of speech to language," *Trends in cognitive sciences*, vol. 4, no. 5, pp. 187–196, 2000.

[8] J. S. Pardo, "Reflections on phonetic convergence: speech perception does not mirror speech production," *Language and Linguistics Compass*, vol. 6, no. 12, pp. 753–767, 2012.

[9] C. Gambi and M. J. Pickering, "Prediction and imitation in speech," *Frontiers in psychology*, vol. 4, p. 340, 2013.

[10] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Attention, Perception, & Psychophysics*, vol. 79, pp. 637–659, 2017.

[11] C. B. Chang, "The phonetics of second language learning and bilingualism," *The Routledge handbook of phonetics*, pp. 427–447, 2019.

[12] D. J. Olson, "Feature acquisition in second language phonetic development: Evidence from phonetic training," *Language Learning*, vol. 69, no. 2, pp. 366–404, 2019.

[13] K. Gnevsheva, A. Szakay, and S. Jansen, "Phonetic convergence across dialect boundaries in first and second language speakers," *Journal of Phonetics*, vol. 89, p. 101110, 2021.

[14] M. A. Wagner, M. Broersma, J. M. McQueen, S. Dhaene, and K. Lemhöfer, "Phonetic convergence to non-native speech: Acoustic and perceptual evidence," *Journal of Phonetics*, vol. 88, p. 101076, 2021.

[15] F. Jiang and S. Kennison, "The impact of l2 english learners' belief about an interlocutor's english proficiency on l2 phonetic accommodation," *Journal of Psycholinguistic Research*, vol. 51, no. 1, pp. 217–234, 2022.

[16] A. J. Olmstead, N. Viswanathan, T. Cowan, and K. Yang, "Phonetic adaptation in interlocutors with mismatched language backgrounds: A case for a phonetic synergy account," *Journal of Phonetics*, vol. 87, p. 101054, 2021.

[17] D. de Jong, A. Pastore, N. Nguyen, and A. D'Ausilio, "Speech imitation skills predict automatic phonetic convergence: a gmm-ubm study on l2," in *Interspeech*, 2022, pp. 769–773.

[18] G. Bailly and A. Lelong, "Speech dominoes and phonetic convergence," in *Interspeech 2010-11th Annual Conference of the International Speech Communication Association*, 2010, pp. 1153–1156.

[19] G. Bailly and A. Martin, "Assessing objective characterizations of phonetic convergence," in *Interspeech 2014-15th Annual Conference of the International Speech Communication Association*, 2014, pp. P–19.

[20] S. Mukherjee, A. d'Ausilio, N. Nguyen, L. Fadiga, and L. Badino, "The relationship between f0 synchrony and speech convergence in dyadic interaction," in *Interspeech 2017*, 2017, pp. 2341–2345.

[21] S. Mukherjee, T. Legou, L. Lancia, P. Hilt, A. Tomassini, L. Fadiga, A. d'Ausilio, L. Badino, and N. Nguyen, "Analyzing vocal tract movements during speech accommodation," in *Interspeech 2018*, 2018, pp. Paper–2084.

[22] S. Mukherjee, L. Badino, P. M. Hilt, A. Tomassini, A. Inuggi, L. Fadiga, N. Nguyen, and A. d'Ausilio, "The neural oscillatory markers of phonetic convergence during verbal interaction," *Human brain mapping*, vol. 40, no. 1, pp. 187–201, 2019.

[23] V. Aubanel and N. Nguyen, "Speaking to a common tune: Between-speaker convergence in voice fundamental frequency in a joint speech production task," *PloS one*, vol. 15, no. 5, p. e0232209, 2020.

[24] M. Garnier, L. Lamalle, and M. Sato, "Neural correlates of phonetic convergence and speech imitation," *Frontiers in psychology*, vol. 4, p. 600, 2013.

[25] M. Babel and D. Bulatov, "The role of fundamental frequency in phonetic accommodation," *Language and speech*, vol. 55, no. 2, pp. 231–248, 2012.

[26] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." in *Interspeech*, 2011, pp. 3081–3084.

[27] V. Delvaux and A. Soquet, "The influence of ambient speech on adult speech productions through unintentional imitation," *Phonetica*, vol. 64, no. 2-3, pp. 145–173, 2007.

[28] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[29] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, 1993.

[30] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[31] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017.

[32] D. de Jong, "Alternating reading task in l2 english," jun 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5588436

[33] Z. Yuan and D. de Jong, "Slovak alternating reading task in l2 english," Jun. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7993783

[34] K. Lemhöfer and M. Broersma, "Introducing lextale: A quick and valid lexical test for advanced learners of english," *Behavior research methods*, vol. 44, pp. 325–343, 2012.

[35] B. McFee, C. Raffel, D. Liang, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. of The 14th Python in Science Conf*, 2015.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[37] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2019.