INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

L1 Identification from L2 Speech Using Neural Spectrogram Analysis

Calbert Graham

Phonetics Laboratory, University of Cambridge, UK
crg29@cam.ac.uk

Abstract

It is well-known that the characteristics of L2 speech are highly influenced by the speakers' L1. The main objective of this study was to uncover discriminative speech features to identify the L1 background of a speaker from their L2 English speech. Traditional phonetic approaches tend to compare speakers based on a pre-selected set of acoustic features, which may not be sufficient to capture all the unique traces of the L1 in the L2 speech for forensic speaker profiling purposes. Convolutional Neural Network (CNN) has the potential to remedy this issue through the automatic processing of the visual spectrogram.

This paper reports a series of CNN classification experiments modelled on spectrogram images. The classification problem consisted of determining whether English speech samples are spoken by a native speaker of English, Japanese, Dutch, French, or Polish. Both phonetically transcribed and untranscribed speech data were used.

Overall, results showed that the CNN achieved a high level of accuracy in identifying the speakers' L1s based on spectrogram pictures without explicit phonetic segmentation. However, the results also showed that training the classifiers on certain combinations of phonetically modelled spectrogram images, which would make features more transparent, can produce results with comparable accuracy rates.

Index Terms: speaker profiling, L1 identification, L2 speech, computational phonetics, forensic phonetics

1. Introduction

Forensic speaker profiling is the analysis of speech to infer the attributes of a speaker. This often involves identifying linguistic features that are associated with particular geographic areas, social groups or unusual pathologies ([1]). Speaker profiling is also relevant in evaluating a claimed identity such as in asylum claims where a wrongful determination could potentially lead to serious consequences. Analysis can also be used to reduce a field of suspects in an investigation by identifying a speaker's background. Characteristics that can be inferred from forensic speaker profiling include sex, age, regional dialect, social background, medical condition, and so on, which may be important to an investigator in narrowing the scope of an investigation.

The present research focusses on the identification of a speaker's native language (L1) from their non-native spoken English. English has become the most widely spoken language globally with the vast majority of its speakers using it as a second language (L2). It is well-known that the characteristic features of these different varieties of English are highly influenced by the native languages of the speakers. Apart from forensic speaker profiling, understanding the speech features that contribute to the perception of foreign-accentedness of a speaker's L2 English may also be useful in other areas such as in foreign language learning (e.g., in pronunciation training systems for the diagnosis and remediation of learner errors).

However, speech analysis for either (forensic or educational) purpose is by no means simple. Many significant challenges exist, including the issue of finding appropriate techniques for modelling the various parameters of the speech signal([2]). In tracing the effects of the L1 on the L2, one of the main difficulties is how to capture the often only subtle differences in how a corresponding phoneme is phonetically realised in the two languages. In modelling L1-L2 interaction, traditional phonetic analyses tend to measure the similarity of an L2 speaker's production (of specific phonemes or prosodic units) as compared to that of a native speaker, based on a pre-selected set of acoustic features. However, apart from being time and expertise consuming, in the context of speaker profiling, the set of extracted features may not be sufficient to capture all the traces of the L1 in the L2 speech that are needed to make an accurate identification of their L1 background.

As an alternative to extracting acoustic parameters, deep learning methods, specifically a convolutional neural network (CNN), can be used to explore the entire visual space of the spectrogram automatically. It is already well-established that articulatory aspects of speech have corresponding specific and unique acoustic correlates ([3]), which are visually represented in the spectrogram as frequency, time and amplitude parameters. However, one main challenge of a CNN-based approach is the lack of phonetic transparency, as it is not evident what specific features are being learnt by the system, making it very difficult to establish specific links to the underlying acoustic details of the spectrogram images. For example, we already know that the acoustic differences between English vowels can be explained in terms of the different locations and widths of the formant frequencies. The frequencies of F1 and F2 – the first two resonances of the vocal tract – largely determine the perceived vowel. A similar relationship also exists between the frequencies of consonants and their articulatory features in voicing, place and manner of articulation such that, for example, F2 points down in labials, toward 1700Hz in alveolars, and pinch with F3 in velar sounds. Phonetic transparency is based on this complex relationship between the way speech is produced and its acoustic representation in a spectrogram image.

Previous research (e.g. [4]) explored various approaches in phonetic modelling that may be used to make deep learning models more interpretable. The present study seeks to apply some of these techniques to the context of native language identification from L2 English speech. More precisely, the main goal of this research is to model L1-L2 interaction to uncover discriminative speech features based on a CNN analysis of spectrogram images that can identify the L1 background of a speaker from their non-native English speech.

The present study addresses three primary research objectives (ROs). RO1 explores whether a CNN approach with

http://dx.doi.org/10.21437/Interspeech.2021-1545

raw spectrogram images as input can learn high-dimensional phonetic features to identify the native language of speakers from their L2 English speech. The hypothesis being tested here is that the spectrogram, which shows the Fourier Transform of a speech signal, contains differences in speech patterns that are formed based on the various ways a speaker's L1 influences their L2 English speech. Previous studies have successfully applied this technique in the area of native language identification (e.g., [5, 6]) with results indicating the CNN networks are able to perform this classification task at high levels of accuracy. In related work, i-vectors have been very effective in modelling speech variability and have also been used successfully as features in classification systems. However, this approach requires significant domain knowledge, making neural networks more appropriate for meaningful feature extraction and classification ([7]).

RO2 is to compare the model in RQ1 with a second model that will be obtained by inputting spectrogram images of vowel and consonant segments extracted from phonetically transcribed data to the CNN classifier. This will allow us to determine which segments are most useful for profiling the L1 of a speaker in the network. A plethora of research has shown that non-native speakers differ systematically from native speakers in the way they produce (and perceive) phonemes in the second language (e.g., [8, 9, 10, 11]). A recurrent observation is that the degree of influence of the L1 on L2 speech depends on the phonological and phonetic similarity or dissimilarity of the two languages and the proficiency of L2 speakers ([8, 12, 13]). If a speaker's L1 and L2 have overlapping phonemes, then finding discriminant features becomes a challenging task for a classifier. RO2 will evaluate the extent to which characteristics of the L1 that are traceable from a speaker's L2 English speech can be used to identify their L1.

RO3 investigates the outcome of applying more detailed phonetic modelling by including information of the transition points between segments in the phonetic model introduced in RO2. The hypothesis being tested is predicated on previous experimental findings showing systematic variability between languages in the spatiotemporal characteristics of coarticulation ([14]) and in how speakers string speech segments together (cf. [15, 16, 17], among many others). This is further supported by previous studies indicating that the relation between production and perception is such that listeners are rigidly attuned to their native language coarticulatory patterns ([18]). RO3 will determine whether the coarticulatory details of how speakers string speech segments together are influenced by the L1 to the extent that it results in unique and detectable acoustic characteristics in their L2 speech.

Taken together, these ROs will allow us to establish the accuracy of identifying the L1 of a speaker from their L2 English in a CNN-based approach with no phonetic modelling (addressed in Experiment 1) and compare it with approaches that include images extracted from segmented speech matching vowels and consonants (addressed in Experiment 2), as well as assess the effects of co-articulatory information (addressed in Experiment 3).

2. Method

2.1. Datasets

Multiple speech corpora were used in this research. We used the IViE corpus ([19]) for native 'Standard Southern British English' (SSBE) and Cambridge English Corpus [29] for non-native English. All English learners identified SSBE as their target variety. On average one minute (randomly selected) speech samples were analysed for each speaker, which were produced in elicited & semi-spontaneous contexts. There were 60 speakers (age range 20-30 years), with an even split for native language (12 SSBE, 12 Dutch, 12 Japanese, 12 French, and 12 Polish speakers) and for gender (30 male, 30 female). L2 speakers were matched at three proficiency levels in each L1 group: 4 advanced, 4 intermediate and 4 basic which corresponded to C1, B2, and A2 on the CEFR scale. The proficiency level of non-native speakers was determined by assessors, who were trained on evaluating oral proficiency according to the CEFR scale. For the relevant experiments (Experiment 2 and Experiment 3, as described below), manual annotations of the data were carried out by two trained phonetic transcribers. Data annotations and processing were performed in Praat ([20]).

2.2. Convolutional Neural Network

The classification problem consisted of determining whether English speech samples were spoken by a native speaker of SSBE, Japanese, Dutch, French or Polish. A CNN was trained as an adapted LeNet ([21]) architecture and Adam Optimizer with batch size 32 in 25 epochs. All analyses were conducted using the Keras package in Python. Input to the model were extracted spectrogram images, as mentioned earlier. The pictures were obtained with a script using the Praat default settings which provided a broadband representation with a 16Khz sampling rate. Dynamic range was set at 70 dB for consistency across speech samples.

2.3. Analysis

Experiment 1 included raw audio signals as input to the CNN network spectrogram images from the spoken English of speakers to be used for L1 classification. Spectrogram images were taken at fixed intervals, with no phonetic transcription. There were 100 images for each of the 60 speakers, giving a total of 6000 images. The training set was 70% (4200 images) and the test set 30% (1800 images). In this and in all subsequent experiments, the Python function model.fit() was used to tell Keras to accept the test set as the validation set.

Experiment 2 was implemented in two parts. In part 2a, the objective was to compare the accuracy of using the spectrogram images of the raw speech signal in Experiment 1 vs. images of phonetically transcribed speech, as already described. Input to this model were spectrogram pictures of (i) vowel segments (specifically, point vowels representing the most extreme F1 and F2 frequencies and for which sufficient data points were available: /i:, u: and ɑ:/) and (ii) fricative and stop consonant segments. Following the methodology in [22], a duration range of 45-250ms for all extracted images was established, with zero-padding to ensure that they were consistently of the same size. Only consonant segments with 20 or more images per speaker were included in the experiment. The CNN was run using images of corresponding transcribed speech segment data, as described before.

Part 2b aimed to elucidate which aspects of the vowel and consonant acoustic properties contain the most meaningful speaker-profiling information. We therefore performed a low-pass filtering (LPF) of target segments (à la [4]). This allows us to determine whether more detailed phonetic modelling can make these findings more interpretable. In the LPF analysis, vowel spectrograms used in Experiment 2a are cropped to obtain 5 versions with maximal frequencies ranging from 250

to 5500 Hz. These images are added to the unmodified sets and are done separately for each vowel. For space reasons only the results for /ɑ:/ are reported in this paper.

Within-word segment transition points contain coarticulatory details that may be the source of speaker-profiling information. In Experiment 3, spectrograms images of inter-segment transition points were added to the images in experiment 2a to evaluate their contribution.

3. Results

3.1. Experiment 1

The results of the experiment with spectrogram images of raw (i.e., untranscribed) audio signals as input indicated that the system was able to learn features to discriminate between the English produced by speakers of different L1 backgrounds on average above 90% accuracy, as shown in Figures 1 and 2. As seen in the confusion matrix of Figure 2, although misclassification rates were relatively low, Dutch L1 was the most misclassified, being identified as English in 6.5% of cases.

Overall, the results confirm the basic assumption of this study that the spectrogram contains sufficient phonetic information to profile the L1 of speakers from their L2 speech. However, as discussed before, to make this link between a speaker's linguistic background and the acoustic production explicit, requires elucidation of how the two relate to the underlying phonetic structure of the speech. This is addressed in Experiment 2.

3.2. Experiment 2

3.2.1. 2a: Transcribed vs Untranscribed data

The CNN was run using images of corresponding transcribed speech segment data, as described before. The results revealed that the network was able to learn features from the transcribed vowel and consonant segments to detect the native languages of speakers at 73% accuracy, albeit at a significantly reduced level of accuracy when compared to the results of the untranscribed model. Results are shown in Figure 5, along with the results of Experiment 3, for convenience.

Further analysis confirmed that among consonants classes voiced continuants (/v, ð, z, ʒ/) yielded the most speaker L1-discriminating information. This finding is depicted in the reduced model shown in Figure 4.

3.2.2. 2b: LPF analysis

The analysis reveals that the lowest accuracy of 51.1% at 250Hz cut-off frequency was significantly above chance (X^2 = 53678, p < 0.01). Further, mid-point value McNemar tests further confirmed that improvement from one model to the next was significant for 250Hz–1000Hz and from 2000–4000Hz (all at p <0.01 level). Model at 1000Hz and at 6000Hz did not show any statistically significant improvement.

3.3. Experiment 3

Overall model performance was at 83.6% accuracy, which is a statistically significant increase of 10.4% on the previous model with full segment images only (X^2 = 73.6, p < 0.01). This confirms that transition points that were hypothesised to contain important coarticulatory details may be a source of L1-discriminating information.

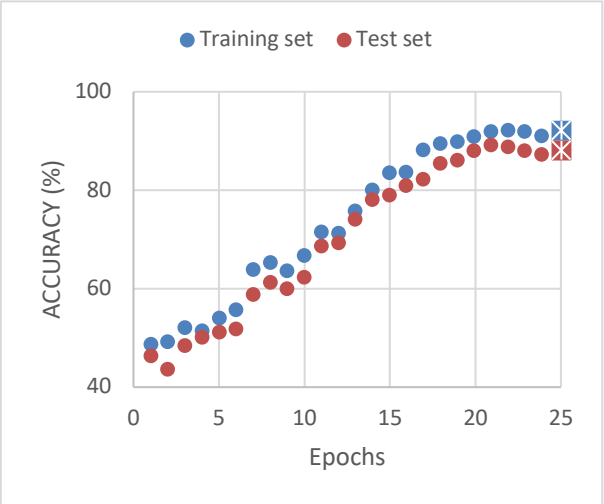

Figure 1: *Overall L1 detection rate.*

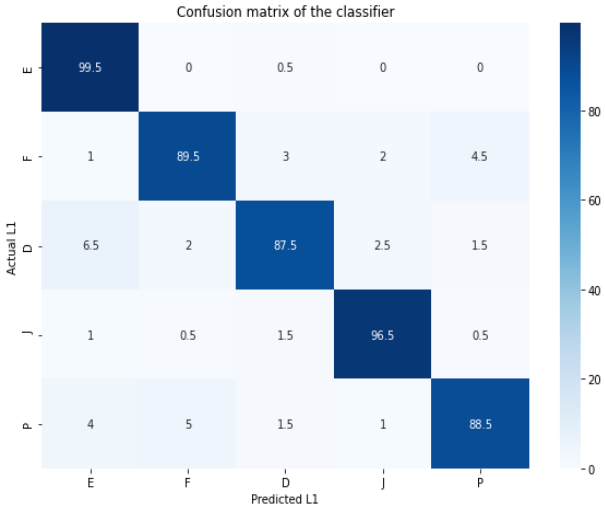

Figure 2: *Confusion matrix of accuracy rates (%).*

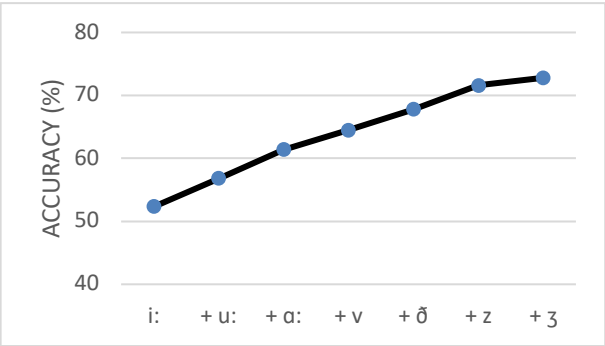

Figure 3: *Accuracy rate for selected segments*

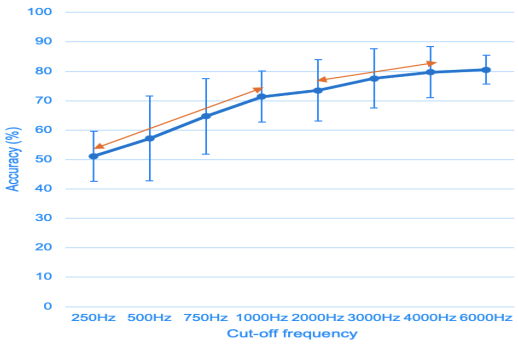

Figure 4: *LPF sample result for the vowel [a:].*

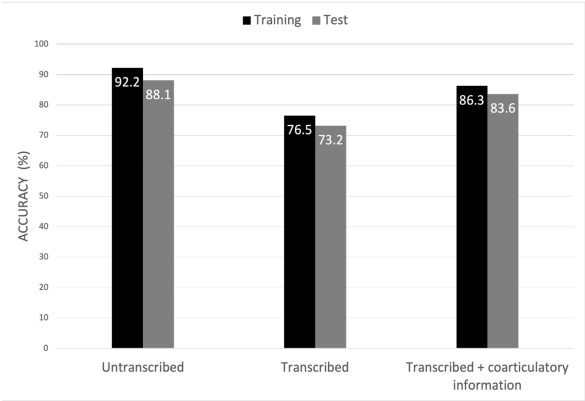

Figure 5: *Overall accuracy rates of all three models*

4. Discussion

As an alternative to the traditional approach of pre-selecting phonetic features for analysis, CNN has the benefit of requiring minimal pre-processing and being able to utilise short audio segments. RO1 investigated the technique in using spectrogram images of raw audio signals as input to a convolutional neural network to perform a multiclass native language identification from L2 English speech. The approach was highly successful in determining the native language of speakers in the experiment. Whilst, overall, there were very small rates of misclassification, the Dutch L1 speakers of English had the highest rate of false negative classifications, with 6.5% of speech samples they produced being misclassified as English. English spoken by native SSBE speakers was consistently the least misclassified variety, suggesting that perhaps the network was performing a thresholding at one layer that separates L1 English from L2 varieties. This would further indicate the possibility that the system had acquired the capability to perform a binary classification of speakers as being either 'native' or 'non-native'. Further, it is remarkable that the system was able to achieve such high levels of accuracy with minimal pre-processing, confirming that this approach is highly accurate and is ideal in contexts where transcriptive data are not available or desirable.

RO2 compared the accuracy of using the spectrogram images of speech in the previous experiment vs. phonetically transcribed speech into vowel and consonant segments. The results indicated that the combination of the three point-vowels with the consonant class of voiced continuants produced the highest recognition rate of 72%. We speculate that the high L1-discriminability among voiced continuants (fricatives) may likely be due to their complex articulatory and acoustic characteristics. Fricatives are generally produced with longer duration than stops and have energy components in the higher frequencies, which may explain why they are often among the last sounds to be acquired by children (e.g., [23, 24, 25, 26]). This further suggests that phonetic modelling can inform deep learning methods in identifying speech segments or segment classes that contain the most L1-disciminating information. Taken together with the results of the LPF analysis, the study confirms that deep learning methods can be effective in identifying speech segments and spectral regions that are crucial to the identification of a speaker's linguistic identity.

Experiment 3 investigated the effect of combining images of the transition points between segments in the CNN model with the full consonant and vowel segments. The results indicated that overall model performance improved at a statistically significant level. This confirms that inter-segmental transition points that contain coarticulatory details are a source of L1-discriminating information. This finding is unsurprising in one sense, given that previous studies have found that coarticulatory information may provide clues to the identity of a speaker ([27]) and may also be the source of language-specific differences in the perception of some sound classes such as fricatives (e.g., [18, 28]). However, the observation that this information has been preserved in the L2 would suggest that previous theoretical approaches that have focused on L1 transfer in segment articulation (e.g., [8]) may be missing an important source of L1-conditioned L2 speech variation by not also considering how speakers string segments together according to the coarticulatory details of their L1. Future research should focus on determining which specific coarticulation features have been used and whether this changes depending on the L1.

5. Conclusions

Overall, the study finds support for the following conclusions: (1) Modelling L1-L2 interactions requires a multidimensional approach to capture that vast variability in speech. (2) Pre-selecting acoustic parameters to extract is both time and expertise consuming and is error-prone, and a deep learning-based approach provides a novel way of accurately modelling L1 transfer through spectrogram image analysis. (3) There is an issue with the transparency of the features learnt in this approach, but research should be focused on making them interpretable. Further, although this study focussed on English (SSBE) as a target L2 variety, the approach is generalisable to other languages and dialects. For the specific task of L1 identification from L2, the results suggest that this approach is highly effective and holds much promise for research in forensic speaker-profiling.

6. Acknowledgements

This work was supported by a Leverhulme Early Career Fellowship awarded to the author. The support of Francis Nolan and Gabriele Chignoli is gratefully acknowledged.

7. References

[1] Watt, D. (2010). The identification of the individual through speech. In Llamas C, Watt D, editors, Language and Identities. Edinburgh: Edinburgh University Press. p. 76-85.

[2] Schilling, N., & Marsters, A. (2015). Unmasking Identity: Speaker Profiling for Forensic Linguistic Purposes. Annual Review of Applied Linguistics, 35, 195-214. doi:10.1017/S0267190514000282

[3] Singh, S., & Singh, K. (1976). Phonetics: Principles and Practices. University park Press. London.

[4] Ferragne, E., Gendrot, C., & Pellegrini, T. (2019). Towards phonetic interpretability in deep learning applied to voice comparison. ICPhS. Melbourne.

[5] Revay, S., & Teschke, M. (2019). Multiclass language identification using deep learning on spectral images of audio signals. arXiv preprint arXiv:1905.04348 (2019)

[6] Sarthak, Shukla, S., & Mittal, G. (2019). "Spoken Language Identification using ConvNets," European Conference on Ambient Intelligence, 252–265.

[7] Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martinez, D., Gonzalez- Rodriguez, J., Moreno, P.: Automatic language identification using deep neural networks. In: 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5337–5341. IEEE (2014)

[8] Flege, J.E.(1995). Second language speech learning theory, findings, and problems. In Winifred Strange(Ed.), Speech perception and linguistic experience: issues in cross-language research (pp. 233–277). Timonium, MD: York Press.

[9] Leather, J., & James, A. (1996). "Second Language Speech". In W. C. Ritchie, T. K., Bhatia (Hgg.), Handbook of Second Language Acquisition. San Diego: Academic Press: 269-316.

[10] Flege, J., MacKay, I.,& Meador, D.(1999). Native Italian speakers' perception and production of English vowels. The Journal of the Acoustical Society of America, 106(5), 2973–2987.

[11] Piske, T., MacKay, I.,& Flege,J. (2001) .Factors affecting degree of foreign accent in an L2: A review. Journal of Phonetics, 29(2), 191–215.

[12] Archibald, J. (1998). Second language phonology, phonetics, and typology. Studies in Second Language Acquisition, 20(2), 189-211.

[13] Graham, C. & Nolan, F. (2019). Articulation rate as a metric in spoken language Assessment. INTERSPEECH: Crossroads of Speech & Language, 17, pp. 3564–3568. Causal Productions.

[14] Manuel, S. (1999). Cross-language studies: Relating language-particular coarticulation patterns to other language-particular facts. In W. Hardcastle and N. Hewlett, editors, Coarticulation: Theory, data and techniques. Cambridge University Press.

[15] Choi, J.D.& Keating, P.(1991) Vowel-to-vowel coarticulation in three Slavic languages, UCLA Working Papers in Phonetics, 78, 78–86.

[16] Boyce, S.E.(1990) Coarticulatory organization for lip rounding in Turkish and English, Journal of the Acoustical Society of America, 88, 2584–2595.

[17] Bradlow, A.R.(1995) A comparative acoustic study of English and Spanish vowels, Journal of the Acoustical Society of America, 97, 1916–1924.

[18] Beddor, P., Harnsberger, J., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. Journal of Phonetics 30, 591–627 doi:10.1006/jpho.2002.0177

[19] Nolan, F., Post, B. (2014). IViE. In Durand, J., Gut, U., Kristoffersen, G. (Eds.), The Oxford handbook of corpus phonology (pp. 475–485). Oxford, UK: Oxford University Press.

[20] Boersma, P. & Weenick, D. (2020). Praat: Doing phonetics by computer, version 5.4.02, http://www.praat.org.

[21] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). "Gradient based learning applied to document recognition," Proceedings of the IEEE.

[22] Guitard-Ivent, F., Chignoli, G., Fougeron, C., & Georgeton, L. (2019). Are IN initial vowels acoustically more distinct. Results from NDA and CNN classifications. Interspeech. Graz.

[23] Dinnsen, D. A. (1996). Context effects in the acquisition of fricatives. In B. Bernhardt, J. Gilbert, & D. Ingram (Eds.), Proceedings of the UBC International Conference on Phonological Acquisition (pp. 136–148). Somerville, MA: Cascadilla Press.

[24] Farwell, C. B. (1976). Some strategies in the early production of fricatives. Papers and Reports on Child Language Development, 12, 97–104.

[25] Ferguson, C. A. (1978). Fricatives in child language acquisition. In V. Honsa & M. Hardman-Bautista (Eds.). Papers on Linguistics and Child Language. The Hague, Netherlands: Mouton.

[26] Moskowitz, B. A. (1975). The acquisition of fricatives: A study in phonetics and phonology. Language, 46, 426–441.

[27] Su, K.-P. Li, and K. S. Fu: Identification of speakers by use of nasal coarticulation Lo-Soun The Journal of the Acoustical Society of America 56, 1876 (1974); doi: 10.1121/1.1903526

[28] Wagner A. (2013). Cross-language similarities and differences in the uptake of place information. J Acoust Soc Am. Jun;133(6):4256-67. doi: 10.1121/1.4802904. PMID: 23742376.

[29] Cambridge English Corpus. Cambridge Assessment/Cambridge University Press. University of Cambridge: https://www.languagesciences.cam.ac.uk/news/cambridge-english-corpus-available-academic-use