INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

End-to-End Acoustic Feedback in Language Learning for Correcting Devoiced French Final-Fricatives

Sucheta Ghosh[1], *Camille Fauth*[2], *Yves Laprie*[1], *Aghilas Sini*[1]

[1]LORIA/CNRS, Nancy, France
[2]Speech Cognition - E.A. 1339, LiLPa, University of Strasbourg, France
sucheta.ghosh@loria.fr, camille.fauth@gmail.com, {yves.laprie,aghilas.sini}@loria.fr

Abstract

This work aims at providing an end-to-end acoustic feedback framework to help learners of French to pronounce voiced fricatives. A classifier ensemble detects voiced/unvoiced utterances, then a correction method is proposed to improve the perception and production of voiced fricatives in a word-final position. Realizations of voiced fricatives contained in French sentences uttered by French and German speakers were analyzed to find out the deviations between the acoustic cues realized by the two groups of speakers. The correction method consists in substituting the erroneous devoiced fricative by TD-PSOLA concatenative synthesis that uses exemplars of voiced fricatives chosen from a French speaker corpus. To achieve a seamless concatenation the energy of the replacement fricative was adjusted with respect to the energy levels of the learner's and French speaker's preceding vowels. Finally, a perception experiment with the corrected stimuli has been carried out with French native speakers to check the appropriateness of the fricative revoicing. The results showed that the proposed revoicing strategy proved to be very efficient and can be used as an acoustic feedback.
Keywords: L2 productions, language learning, speech signal processing, acoustic feedback, speech perception, voicing.

1. Introduction

This work aims at providing acoustic feedback to help learners of French to pronounce voiced fricatives. Like other EU languages (except Basque, Norwegian, and Saami), French has voiced ([v,z,ʒ]) and its corresponding voiceless ([f,s,ʃ]) fricatives [1], though the voicing profiles are different [2, 3], for instance, French has no final devoicing but German has[4]. Indeed, the Germanic languages are more prone to devoicing than the Romance languages [2, 5, 6]. German learners face hardest difficulty to produce French voiced fricatives in word-final position [7]. In fact, voiced fricatives are more prone to devoicing than comparable stops[8]. Indeed, two pressure drops are needed to produce voice fricatives: first at the glottis to enable vocal fold vibration, then across the constriction to generate a turbulence downstream and consequently a noise sound [9], therefore all else unchanged, the production of voiced fricatives is a more complex task than that of their voiceless counterparts.

Therefore, to facilitate the perception and production of fricatives of the learners, we propose a novel framework that incorporates a correction method which substitutes the erroneous devoiced fricative with voiced one using TD-PSOLA [10], and also the energy of the replacement fricative was adjusted w.r.t. the preceding vowels of the learner and teacher. Before this, the framework automatically detects learners' voiced/unvoiced word utterances, and it also selects good exemplars from a French native corpus with respect to the learner utterance. A perception experiment successfully validates the outputs of this framework. This approach is applicable to similar cases where concatenative synthesis can be used, for example the case of final voiced stops.

The proposed framework utilizes a range of acoustic cues. The fricatives are basically characterized by the spectral properties of the friction noise, amplitude and duration of the noise, and formant trajectories or spectral properties of the transition with the surrounding vowels [11]. In this work we consider voiced fricatives in the same vocalic context, that is, /a/, to share the same formant trajectories at the boundary between the vowel and fricative [12, 13]. The voicing during the fricative segment is found as the dominant cue for the listeners' voicing decision[14], though the scope of this paper is beyond the debate of incompleteness of voicing neutralization [15]. The fricative duration cue plays a significant perceptual role to distinguish between voiced and voiceless fricatives, and also lengthening the duration of the preceding vowel of the word-final fricative is an effective technique to produce a voiced fricative [16, 10]. Low-frequency energy was included as a successful measure of voicing during the frication by a series of works [17, 18]; here a spectrum over the whole frication noise was computed, and then the amplitude of the components below 500 Hz was measured.

In the next section we describe and analyze the corpus data used in this work, then we detail about the proposed framework and its output analysis, then we report decisive results of a perception experiment with those corrected stimuli and conclude.

2. Corpus

2.1. IFCASL Corpus & Analysis

In this work we used words containing voiced fricatives at final position, extracted from the sentences of bilingual IFCASL corpus [19, 20]. The corpus is recorded by French learners of German and German learners of French in their native and L2. Here we use two groups: the native German learners of French and the native French. Approximately 40 speakers from each of the two groups recorded the corpus in a quiet room reading aloud sentences displayed on a laptop. The 16-bit audio data was recorded at a sampling frequency of 22 kHz. The gain was automatically controlled during the recording session.

Six sentences, containing one of the three French voiced fricatives in a word-final position (and located at the end of an accentuated group) were gathered from whole corpus; each of these voiced fricatives occurs in two sentences: for /ʒ/ 1. "Elle habite dans un beau **village** en France."(/vilaʒ/) 2. "Mon ami a perdu ses **bagages** à la gare."(/bagaʒ/); for /v/ 1. "Le champagne est rangé dans la **cave** en terre."(/kav/) 2. "Les enfants sont **braves** à l'ecole."(/brav/); for /z/ 1. "Les avions sont rentrés à la **base** après le vol."(/baz/) 2. "Les élèves doivent cocher la bonne **case** avec un feutre."(/kaz/). These bold-faced words are

http://dx.doi.org/10.21437/Interspeech.2017-1031

Table 1: *A comparison of the fricative-segment features. Italics font denotes significant feature at 95% test level.*

Fricative-Properties			**FF voiced**				**G0 devoiced**	
	v	z	ʒ	Overall	v	z	ʒ	Overall
Voiced % of Frames	97.4	84.2	92.0	**91.2**	16.2	15.4	20.0	**17.2**
Energy/vowel(dB)	-11	-4.46	-0.88	**-5.3**	-3.64	-1.49	4.79	**-0.11**
Low Freq. Energy(dB)	69.16	67.24	68.63	**68.34**	57.20	51.49	59.90	**56.20**
Spectrum Energy(dB)	47.77	53.99	57.02	**52.93**	51.31	53.28	59.1	**54.56**
Duration								
% w.r.t. vowel	46.4	43.2	44.7	**44.7**	74.8	94.6	60.2	**76.5**
in MS	49	55	53	**52**	93	137	101	**110**

extracted from the sentence utterances. Of all these word utterances, a set of 111 devoiced utterances by German natives (say, G0) are found by two phoneticians, judging the voiced feature of the fricatives as voiced or unvoiced by using the spectrogram, the $F0$ values, and also by listening to the vowel-fricative sequences (that is, [bagaʒ]:15, [brav]: 25, [kaz]: 24, [vilaʒ]: 14, [kav]: 10 and [baz]: 23). We randomly chose another set of 111 (number breakdown identical to G0) French native word utterances, say FF, of IFCASL to use them in our analyses and experiments.

The corpus analysis of acoustic cues realized by the groups G0 and FF allowed us to identify the difficulties faced by learners and thus to design acoustic feedback. In Table. 1 the computed values of the percentage of voiced frames (non-zero $F0$ values), the energy of the fricative with respect to that of the vowel considering the 0-8 kHz frequency band, the low-frequency energy (under 500 Hz), the spectrum energy for 0-8 kHz frequency band, the absolute duration of the fricative and the ratio between the fricative and the vowel duration, are shown. The unpaired t-test with unequal variances for all these features revealed that the differences for $F0\%$ ($p < 0.001$), low-freq. energy ($p < 0.01$), fric/vowel (that is, fricative/vowel) duration ($p < 0.05$), and absolute duration ($p < 0.01$) features of FF and G0 groups are significant (at 95% test level), other features were not found statistically significant for those differences[21]. As expected, the German speakers achieved much lower amount of voiced frames, and produced much longer fricatives in absolute values and also with respect to the previous vowel; moreover they produced more energetic fricatives than the native French speakers.

As the acoustic cues identified by the experts are tightly related to the voicing, so it is required to provide learners with acoustic feedback that corrects relevant acoustic cues. There are several attempts to provide acoustic feedback using the users' own voice [22, 23], mostly using PSOLA techniques for prosodic modifications. These techniques can also provide spectral transformations performed through DTW training [24] or by using an algorithm capable of modifying the LPC envelope [25]. Here we focused on the TD-PSOLA synthesis technique that is more appropriate than other analysis/synthesis algorithms such as the Harmonic/Stochastic (H/S) model [26] or multi-band resynthesis MBR-PSOLA [27]. Adding synthetic voicing via the use of a source model raises issues to ensure a continuous phase at the boundary of the vowel and the new voiced signal in low frequency. Therefore, we decided to replace the unvoiced fricative uttered by the German speaker with a voiced fricative uttered by a French speaker. We resorted to concatenative synthesis via PSOLA to implement the replacement. Using voiced fricatives pronounced by the French speakers of the IFCASL corpus could mislead learners because the influence of the next word, even if it is at the end of an accented group, could deviate from the learner's realization. Also, in this work we focus on the acoustic feedback of isolated word with the devoiced fricative, which requires well realized words with a voiced fricative. Consequently, we recorded the same six words as a corpus of isolated words for this modification purpose.

2.2. French Isolated Fricative Word Corpus

Three repetitions of the six words from IFCASL (with a final voiced fricative) were recorded by 45 French natives (Male: 22, Female: 23; Speaker-age: 20-58). Carrier sentences were not used to avoid risk to alter the articulation of the isolated words at their extremities. The finest exemplars (that is, those which do not exhibit strong $F0$ effects and with a correct fricative duration) are selected to use in correction to avoid list-effect in production of isolated words. We found such 140 isolated word utterances ([bagaʒ]: 11, [vilaʒ]: 13, [kav]: 31, [brav]: 31,[baz]: 28, [kaz]: 26). Since our study involves the fricatives at word-final position where schwa is also frequently found. From a production point of view the realization of a voiced fricative requires an abduction/adduction movement of the vocal folds corresponding to the apparition/disappearance of a frication noise. But, this movement has to be well calibrated to prevent devoicing. The adjustment of this movement seems all the easier since there is a vowel after the fricative. This may explain why French speakers often realize a vocalic release, that is, a vocal schwa after a voiced fricative[28]. This final schwa does not modify the identification of the word and is neutral for French listeners' point of view. These corpora and the results of analysis are used in the proposed end-to-end framework in the next section.

3. End-to-End Framework for Acoustic Feedback

Given a learner's fricative word utterance to the framework, first it pass through **(i) Voiced/unvoiced Utterance Detection:** the voiced/unvoiced utterances are classified using a hard majority-voting based ensemble classifier[29]. Three different individual classification models classify the samples: Logistic regression, a naive Bayes classifier with a Gaussian kernel, and a random forest classifier, then these three models are created as ensemble to balance out individual weakness of the classifiers. We use the significant features from the Table.1 in this classification using merged FF and G0 datasets. The accuracy of the classifier is 0.98 with 10-fold cross-validation of data (standard deviation: 0.02) on training data (80% of full dataset), whereas the test (rest 20% of data) accuracy is 0.96. If the input utterance is found unvoiced then it passes through **(ii) Teacher Utterance Selection w.r.t. Learner's:** in order to carry out an intelligible correction sounding natural, one voiced fricative has to be selected from the database of isolated words (sec. 2.2). We chose one that gives the minimal distance between the $F0$ curves of the learner and the teacher[30]. Then the teacher's voiced and learner's unvoiced utterances pass through **(iii) Re-voicing via PSOLA:** TD-PSOLA is a simple and efficient technique for implementing concatenative synthesis[31]. However, concatenation itself should be realized carefully to prevent any acoustic discontinuity and acoustic artifacts. Usually, concatenation is realized between diphones at phone centers. Connecting diphones at the center of vowels (here /a/) is the least favorable case because connecting periods without introducing any phase shift is difficult and more likely to generate an audible artifact at this point. Furthermore, the concatenation algorithm usually explores a huge corpus to find out a sequence of units that minimizes a concatenation and target cost. In our case the choice is more constrained as the vowel before the fricative uttered by the learner, cannot be changed. Additionally, unlike text-to-speech synthesis that exploits one speaker corpus, in this case the substituted voiced

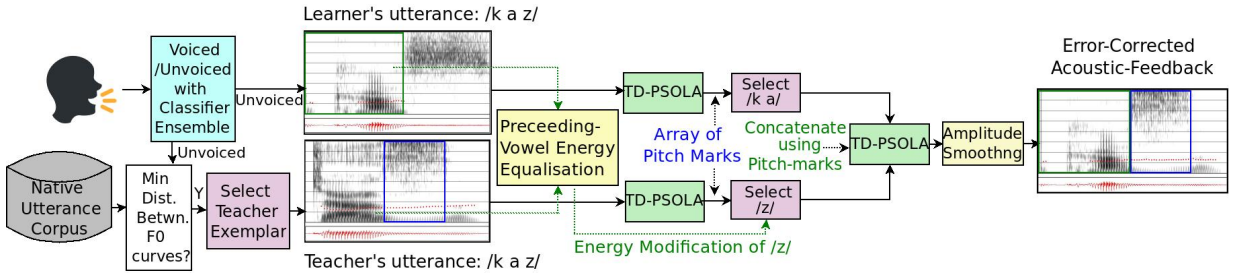

Figure 1: *End-to-end process-flow for feedback correction with spectrograms of /k a z/ with unvoiced /z/ (learner) and after re-voicing of /z/. The original F0 curves The red dotted lines superimposed on the spectrogram represent the original and final $F0$ curves.*

fricative has not been pronounced by the same speaker as there is no guarantee of the quality of the fricatives.

Hence we opted for another concatenation strategy which also exploits the robustness of the segmentation and efficient forced alignment provided by automatic speech recognition (see Fig. 1). In fact, the boundary between a vowel and an devoiced fricative can be found with a very good precision by ASR since these sounds are spectrally very different. Besides, acoustic models used by ASR were trained by incorporating non-native data into a French native corpus to partly overcome the problem of acoustic deviations due to the non-native accent [32]. Thus, we chose to concatenate the replacement fricative at the boundary.

To obtain a seamless energy concatenation we passed through a process of comparative equalization of energy of teacher's and learner's fricative segments with respect to the preceding vowel energy, since the energies of the learner's devoiced and the teacher's voiced fricatives are intrinsically different (see Fig. 1 learner & teacher signal). First the energy difference between the learner's and the teacher's vowels is computed, then the teacher's fricative segment is multiplied by factor $P(E)$ such that:

$$P(E) = \exp\left(\frac{\Delta E}{20} \times ln(10)\right)$$

where,

$$\Delta E = \frac{\sum_{j=0}^{vD_{min}} \sum_{i=0}^{f_{max}} vE_l[i,j] - \sum_{j=0}^{vD_{min}} \sum_{i=0}^{f_{max}} vE_t[i,j]}{(vD_{min} * f_{max})}$$

where, $vD_{min} = \min\{vDur_t, vDur_l\}$, $vDur_l$ is the duration of the preceding vowel segment of learner, $vDur_t$ is the duration of the preceding vowel segment of teacher; $f_{max} = 2$ kHz, that is the range of analysis for the spectrogram is from 0 to 2 kHz; vE_l and vE_t represent the (frame-wise) energy of vowel segments for the learner and teacher respectively.

We already noted that several French isolated word utterances comprise of a schwa after the voiced fricative at the end; this is also well recognized and segmented by the ASR, so we made an attempt to correct each of the voiceless utterances twice, that is, the correction of fricatives with a schwa (say, G+) and the correction of fricatives without schwa (say, G-). Although the additional schwa used in the G+ condition is perceptually transparent, but probably less perceptually transparent for Germans, it probably could help the speaker to adopt the same production strategy, which keeps the abduction/adduction movement of glottis at a level that allows vocal fold vibration.

3.1. Analysis of Fricative Energy after Correction

Since the energy has a prominent impact on the perception of the voicing feature we performed energy analysis in two ways: **(a)** the difference in dB (called energy/vowel in Fig. 2a histogram) between the energy of the fricative and that of the preceding

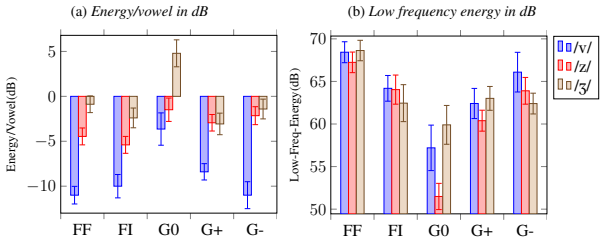

Figure 2: *Histograms of energy related properties of fricative durations (error bars show 95% confidence level), for FF: French IFCASL voiced utterances, FI: French Isolated utterances, G0: German IFCASL voiceless utterances, G+&G-: Corrected G0 utterances respectively with&w/o schwa for 3 types of fricatives.*

vowel (both energies are computed within the range 0-8 kHz and averaged over the duration); **(b)** low freq. energy (below 500 Hz following [17]) for the five groups per word basis (shown in the histogram of Fig. 2b). From Fig. 2 we observe that: **(i)** in case of low freq. energy, the energy values of devoiced G0 group are the lowest (paired t-test G+ & G0: $p < 0.1$, ns at 95% significance-level, but significant at 90%-level, G- & G0 paired p = 0.11 ns) **(ii)** in case of energy/vowel (dB), the ratio-values of devoiced G0 group are highest (paired t-test G+ & G0 paired $p < 0.1$, ns at 95% -level, but significant at 90%-level, for G- & G0, p = 0.14 ns at 95% level) **(iii)** after correction both the low freq. energy and energy/vowel values of the G- and G+ groups are comparable to those of the voiced groups FF (IFCASL French native) and FI (French native Isolated). **(iv)** the overall confidence interval of G+ group is narrower, thus stronger than G- group.

4. Perceptual validation of the acoustic feedback

Since the feedback is proposed to the learners with an intention to guide them towards a correct articulation, the correctness of these should be judged by native French speakers. Thus, this perception experiment aims at evaluating how well the voiced feature is perceived after having the applied feedback.

In this experiment we used the full four sets of data (the same number of stimuli per group), namely, French native isolated words extracted from IFCASL (FF), devoiced word utterances by German natives from IFCASL (G0), corrected (re-voiced) word utterances of G0 without schwa (G-) and the same with schwa (G+). In order to stipulate the duration of the experiment we randomly selected 402 stimuli out of all the 444 stimuli for each of the listeners. All the fifty French native participant listeners were rewarded. We used the full words as stimuli, and specifically instructed the listeners to focus on the word-final sounds only, not any other part of the utterances. We made the voiced fricative rating choice more flexible with a 5-point scale; additionally, we kept another question in 5-point scale on the

possible belonging to the L1 or L2 groups, for the confidence measurement, then we binarize the ordinal data following [33].

4.1. Results & Discussions

Fig. 3 and 4 depict the identification accuracy rates of the listeners with and without the expert opinion respectively. Fig. 4 gives the identification accuracy percentage only on the basis of the listeners' majority voting [34] for each item of four groups. A rate of 100% for the G0 condition would mean that all the stimuli identified as devoiced by the experts are also perceived as devoiced by listeners, whereas a rate of 100% for the FF, G+ and G- conditions means that all the stimuli are perceived as voiced.

If the re-voicing were perfect we would expect the auditory judgment to switch from one extreme to the other, that is, from completely unvoiced for G0 to completely voiced for G- and G+ with an accuracy of 100%. Fig. 3 shows that the voicing judgment almost completely switched from unvoiced to voiced after correction since 91% of the G+ stimuli and 89% of the G- are judged voiced as expected. Furthermore, the identification rate is very close to the rate obtained with stimuli uttered by French speakers (FF). The statistical significance (at 95% test level) of the results, computed through paired t-test for the effect of correction, fully validate our re-voicing strategy [21].

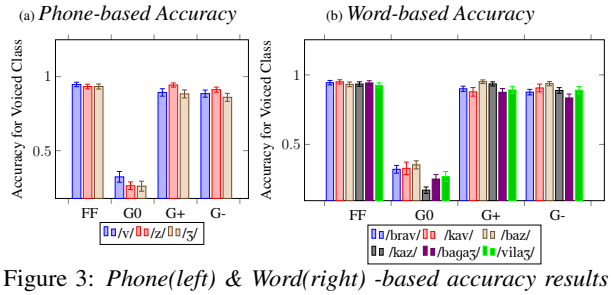

Figure 3: *Phone(left) & Word(right) -based accuracy results of perception of corrected stimuli (error bars show confidence intervals at 95% level) for FF: French IFCASL voiced utterances, G0: German IFCASL devoiced utterances, G-&G+: Corrected G0 utterances w/o & with schwa respectively.*

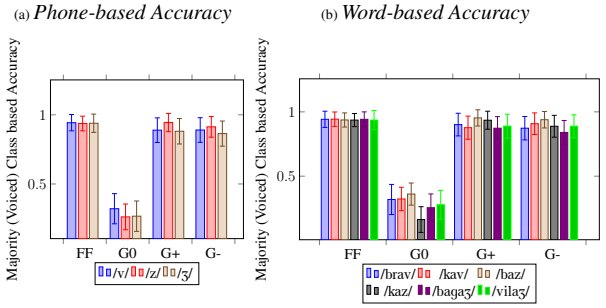

Figure 4: *Phone(left) & Word(right) -based accuracy results of voiced perception of corrected stimuli, without using data by experts, using majority voting (error bars show confidence interval at 95% level) for FF: French IFCASL voiced utterances, G0: German IFCASL devoiced utterances, G-&G+: Corrected G0 utterances w/o & with schwa respectively.*

A detailed review of the Fig 3a reveals a very good significant difference of correction for three fricative categories (G0 & G+:$p < 0.01$; G0 and G-: $p < 0.01$). The detailed results of Fig. 3b show that the correction performed equally well for the six fricative words, which are approximately identified at the identical rate, though for the G0 condition it demonstrates a lower unvoiced identification rate for /v/ than for /ʒ/ before correction. Overall results with six words show a consistent and strong statistically significant difference effects of corrections (G- vs. G0: $p < 0.001$; G+ vs. G0: $p < 0.001$). We see that the confidence intervals of the G- and G+ conditions are generally closer to that of the FF condition. This shows that the correction allows an easy identification and a good agreement between listeners, whereas the confidence interval of the G0 is wider than the FF, G+ and G- cases, due to the higher uncertainties for listeners' decisions for the devoiced G0 cases.

Finally, the condition G+ with schwa gives slightly better results than the G- condition. This means that the vocalic release is recognized as a natural feature by French listeners and has a prominent impact on phone voicing identification (G+ & G-: $p < 0.1$, not significant at 95% significance level, but significant at 90%-level). From Fig. 4 it can be seen that the results are very close to those obtained with the experts' judgment. These results completely corroborate the efficiency of the re-voicing strategy. The result differences based on the six words are statistically significant for the effect of corrections (G+ vs. G0: $p < 0.001$; G- vs. G0: $p < 0.002$). In this case as well, G+ performs slightly (though non-significant) better than the G- condition.

5. Conclusion

The phonetic analysis of the realizations of French final voiced fricatives shows the nature of difficulties faced by German speakers to produce the expected voicing feature. Our primary analysis results also confirm that the voicing feature incorporates several acoustic correlates, on the levels of voiced percentage of $F0$ frames of fricative duration, the ratio of fricative and vowel duration and fricative energy with respect to the preceding vowel. In this work we propose an end-to-end framework for acoustic feedback to facilitate the perception and production of French final voiced fricatives. We essentially used a TD-PSOLA method to design the acoustic corrections, together with the fricative energy correction of the acoustic feedback, w.r.t. the preceding vowels of the learner and the exemplar. To limit interactions with acoustic features of other speech segments, the current corrections focus on simple utterances to ensure the reliability of phone segmentation via ASR, and consequently the relevancy of the corrections. Besides, in this framework we also used a classifier ensemble to detect the voiced/unvoiced utterances, and we chose a best fitting exemplar for the learner utterance using minimum F0 curve distance procedure. Finally, a perception experiment with corrected stimuli reveals that the concatenation strategy worked very efficiently. The benefit for German learners of French, and especially the comparison between the G- and G+ conditions and the long term effect will be investigated in a near future. The proposed method goes far beyond the case of the word-final devoiced fricatives by German learners of French. It is applicable to all similar cases where a concatenative synthesis can be used without introducing audible acoustic artifact. The case of final voiced stops is very similar from a processing point of view since there is a well marked transition between the vowel and the plosive. This situation lends itself well to the same strategy of re-voicing via PSOLA concatenation exploiting a base of correctly pronounced voiced stops.

6. Acknowledgements

This work is supported by an ANR/DFG Grant "IFCASL" to the Speech Group LORIA CNRS UMR7503 Nancy France and to the Phonetics Group, Saarland University Saarbrücken Germany.

7. References

[1] I. Maddieson, "WALS - Voicing in Plosives and Fricatives," 2010. [Online]. Available: http://wals.info/feature/description/4

[2] L. M. Jesus and C. H. Shadle, "A parametric study of the spectral characteristics of European Portuguese fricatives," *Journal of Phonetics*, vol. 30, no. 3, pp. 437–464, 2002.

[3] J. Sieczkowska, B. Möbius, and G. Dogil, "Specification in context-devoicing processes in Polish, French, american English and German sonorants," in *INTERSPEECH. 2010*, 2010.

[4] B. Möbius, "Corpus-based investigations on the phonetics of consonant voicing," *Folia Linguistica*, vol. 38, no. 1-2, pp. 5–26, 2004.

[5] D. Pape, L. Jesus, and P. Jackson, "Devoicing of phonologically voiced obstruents: Is European Portuguese different from other Romance languages," in *the 17th International Congress of Phonetic Sciences*, 2011, pp. 1566–1569.

[6] D. Pape and L. M. Jesus, "Production and perception of velar stop (de) voicing in European Portuguese and Italian," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 1, p. 1, 2014.

[7] D. Jouvet, A. Bonneau, J. Trouvain, F. Zimmerer, Y. Laprie, and B. Möbius, "Analysis of phone confusion matrices in a manually annotated French-German learner corpus," in *Workshop on Speech and Language Technology in Education*, 2015.

[8] J. Ohala, "Aerodynamics of phonology," in *4th Seoul International Conf. on Linguistics [SICOL]*, 1997, pp. 92–97.

[9] K. Stevens, "Airflow and turbulence noise for fricative and stop consonants: static considerations," *The journal of the acoustical society of America*, vol. 50:4, no. 2, pp. 1180–1192, 1971.

[10] S. Ghosh, C. Fauth, A. Sini, and Y. Laprie, "L1-L2 Interference: The case of final devoicing of French voiced fricatives in final position by German learners," in *Interspeech 2016*, 2016, pp. 3156–3160.

[11] H. Reetz and A. Jongman, *Phonetics: Transcription, Production, Acoustics, and Perception*, 1st ed. Wiley-Blackwell, 2011.

[12] K. Stevens, S. Blumstein, L. Glicksman, M. Burton, and K. Kurowski, "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," *Journal of the Acoustical Society of America*, vol. 91, no. 5, pp. 2979–3000, 1992.

[13] L. F. Wilde, "Analysis and synthesis of fricative consonants," Ph.D. dissertation, Massachusetts Institute of Technology, 1995.

[14] D. Pape, L. M. Jesus, and P. Birkholz, "Intervocalic fricative perception in European Portuguese: An articulatory synthesis study," *Speech Communication*, vol. 74, pp. 93–103, 2015.

[15] O. Dmitrieva, A. Jongman, and J. Sereno, "Phonological neutralization by native and non-native speakers: The case of Russian final devoicing," *Journal of Phonetics*, vol. 38, no. 3, pp. 483–492, Jul. 2010.

[16] D. H. Klatt, "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America*, vol. 59, no. 5, pp. 1208–1221, 1976.

[17] B. McMurray and A. Jongman, "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," *Psychological Review*, vol. 118, no. 2, pp. 219–246, 2011.

[18] L. Spinu and L. Jason, "A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives," *Journal of Phonetics*, vol. 57, pp. 40–58, 2016.

[19] J. Trouvain, B. Andreeva, V. Colotte, C. Fauth, D. Fohr, D. Jouvet, J. Jügler, Y. Laprie, O. Mella, B. Möbius, and F. Zimmerer, "The IFCASL corpus of French and German non-native and native read speech," in *10th Language Resources and Evaluation Conference (LREC)*, 2016.

[20] C. Fauth, A. Bonneau, F. Zimmerer, J. Trouvain, B. Andreeva, V. Colotte, D. Fohr, D. Jouvet, J. Jügler, Y. Laprie, O. Mella, and B. Möbius, "Designing a bilingual speech corpus for French and German language learners: a two-step process," in *Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 1477–1482.

[21] R-Core-Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, 2013. [Online]. Available: http://www.R-project.org

[22] K. Hirose, F. Gendrin, and N. Minematsu, "A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice," in *INTERSPEECH. 2003*, 2003.

[23] M. Pfitzinger and H. Bissiri, "Italian speakers learn lexical stress of German morphologically complex words," *Speech Communication*, vol. 51, pp. 933–947, 2009.

[24] H. Valbret, E. Moulines, and J.-P. Tubach, "Voice transformation using PSOLA technique," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 145–148.

[25] F. G. de Los Galanes, M. H. Savoji, and J. M. Pardo, "New algorithm for spectral smoothing and envelope modification for LP-PSOLA synthesis," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference*, 1994.

[26] Y. Stylianou, "Harmonic plus Noise Models for Speech combined with Statistical Methods for Speech and Speaker Modification," Ph.D. dissertation, École Nationale Supérieure des Télécommunications, 1996.

[27] T. Dutoit and H. Leich, "MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, pp. 435–440, 1993.

[28] Calliope, "Description acoustique," in *La parole et son traitement automatique*. Paris: Masson, 1989, ch. 3.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] R. A. Clark and K. E. Dusterhoff, "Objective methods for evaluating synthetic intonation," in *Proceedings of EUROSPEECH*, 1999.

[31] E. Charpentier and F. Moulines, "Pitch synchronous waveform processing techniques for a text-to-speech synthesis using di- phones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[32] D. Jouvet, H. Mesbahi, A. Bonneau, D. Fohr, I. Illina, and Y. Laprie, "Impact of pronunciation variant frequency on automatic non-native speech segmentation," in *Language and Technology Conference - LTC'11*, 2011, pp. 145–148.

[33] G. Norman, "Likert scales, levels of measurement and the "laws" of statistics," *Advances in health sciences education*, vol. 15, no. 3, pp. 625–632, 2010.

[34] L. Breiman, "Pasting small votes for classification in large databases and on-line," *Machine Learning*, vol. 36, no. 1, pp. 85–103, 1999.