INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

L1-L2 interference: the case of final devoicing of French voiced fricatives in final position by German learners

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

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

Abstract

This work is dealing with a case of L1-L2 interference in language learning. The Germans learning French as a second language frequently produce unvoiced fricatives in word-final position instead of the expected voiced fricatives. We investigated the production of French fricatives for 16 non-native (8 beginner- and 8 advanced-learners) and 8 native speakers, and designed auditory feedback to help them realize the right voicing feature. The productions of all speakers were categorized either as voiced or unvoiced by experts. The same fricatives were also evaluated by non-experts in a perception experiment targeting VCs. We compare the ratings by experts and non-experts with the feature-based analysis. The ratio of locally unvoiced frames in the consonantal segment and also the ratio between consonantal duration and V1 duration were measured. The acoustic cues of neighboring sounds and pitch-based features play a significant role in the voicing judgment. As expected, we found that beginners face more difficulties to produce voiced fricatives than advanced learners. Also, the production becomes easier for the learners, especially for the beginners, if they practice repetition after a native speaker. We use these findings to design and develop feedback via speech analysis/synthesis technique TD-PSOLA using the learner's own voice.
Index Terms: L2 productions, language learning, speech signal processing, acoustic feedback, speech perception.

1. Introduction

It can be taken for granted that the first language (L1) influences the target language to be learned (L2) on all linguistic levels including the lexicon, morphosyntax, pragmatics and, most pertinent to our objective, the sound structure and its phonetic implementation (e.g. [1] or [2]). The phonetic realization of the phonological distinction between voiced and unvoiced consonants can show a huge variability between different languages, often in dependence of the position in the syllable and the word as shown by the cross-linguistic studies by [3] and [4]. It is found that the Germen learners face most of the difficulties to pronounce the French voiced fricatives [5]. Alongside, the occurrence of occlusives in the same word worsens the situation. This study deals with one problem of L1-L2 interference on the segmental level: the realization of fricatives in word-final position (which are subject to devoicing in German, but not in French [6]) and the identification of the acoustic features which support the voiced/unvoiced judgment by native speakers. Now the question is, what will be the method of formulation of the feedback by means of the knowledge we acquire from our study? From a learner's perspective, after accepting the need to change pronunciation habits, the L2 speech learning process [7, 8] starts with raising the awareness of the distinctions to be learned. Only a few attempts of giving feedback in the learner's own voice have been reported, but these appear to be quite effective [9, 10]. The first part of the paper is intended to assess the learners' realizations of word-final voiced fricatives by measuring acoustic cues linked to voicing. The second part (Sec. 6) is devoted to the correction strategies, i.e. the re-voicing of devoiced fricatives produced by German speakers learning French. The novelty of this paper resides in the choice of the study of the language pair for L1-L2 interference, and as well in the proposed method of feedback.

2. Methods

2.1. Speakers

In this work, we use the utterances of 16 German native speakers learning (L2) French: 8 beginners (A2 level according to the CEFR [11]) and 8 advanced (C1 level) learners, and also 8 French native speakers as control speakers from IFCASL corpus [12]. For acquisition, we used Corpus Recorder software [13]. The output is a set of wav format files that stores the information about speaker, mother tongue, level, gender, sentence. The recordings were made in a quiet room, without reverberations. A headset microphone (AKG C520) and an audiobox (M Audio Fast track), plugged into a Windows laptop, were used as recording devices. All sessions were carefully monitored to ensure consistent quality.

2.2. Corpus

We focus only on the production of French obstruents, specifically [b,d,g,v,z,ʒ], of IFCASL corpus [12]. Three groups, namely French native speakers (Control Speakers or CS), German-A2 level learners (Beg) and German-C1 level learners (Adv), took part in the experiments as speakers. In this task, we selected 9 sentences divided into two different tasks. In the first task, say SR (Read-Only task), the speakers are asked to read aloud the sentence (3 different sentences). In the second, say SH (Read-as-Repetition task), the learners listen to the sentence one or several times, pronounced by a French native speaker, then they are instructed to read that same sentence aloud (6 different sentences). To investigate the realization of voiced fricatives in the final position, we selected 3 words from the SR task and 3 words from the SH task in the corpus. The words contain voiced fricatives. The obstruent is in final-syllabe position preceded by the same vocalic environment [a] and followed by a vowel. Specifically, we chose V1CV2 (Vowel1-Consonant-Vowel2) pattern (e.g. "Les avions sont rentrés à la **base** après le vol."; "Les élèves doivent cocher la bonne **case** avec un feutre.").

http://dx.doi.org/10.21437/Interspeech.2016-954

3. Hypothesis

We study to what extent the German learners are able to produce voiced fricatives in the word-final position, and test whether their production: 1) depends on the speaker's level: maybe L2 beginner speakers (A2 level) have more difficulties to produce a voiced fricative [v,z,ʒ] in the final position than C1 speakers. 2) depends on the task: production could be easier in the repetition task than in the read-only condition.

4. Measurements

First, the corpus was automatically segmented and annotated by the speech to text alignment tool [14, 15]. Then, the corpus was checked with respect to the orthographic transcription. It was manually checked at the levels of phones and words (phonetic transcription) and corrections were made if necessary. It is to be noted that if the consonant is partially devoiced in some cases, this does not necessarily lead to the perception of the corresponding unvoiced consonant. We measured the fraction of locally unvoiced frames in the consonantal segment (in %). If there is no suitable pitch candidate in a frame, the frame is considered voiceless [16]. We also calculated the ratio of consonantal duration to V1 duration because according to the literature the consonant seems to be longer when it is devoiced (see for example [3]). Thus, we designed an intuitive feature set comprising seven features for each of the two settings, namely, SR (read-only condition) and SH (read-as-repetitions condition). The seven features are: 1. the fraction of voiced frames to the number of frames of the consonant in % (F0Percentage) 2. Zero Crossings Rate (ZCR) 3. Word duration (W) in milliseconds (ms) 4. Vowel1 duration in ms (V1), 5. Fricative duration in ms (C) 6. ratio of the fricative duration to the word duration (C/W) and 7. ratio of fricative duration to vowel1 duration (C/V1). Due to the lack of space we only show results for the representative words: "case" for SR setting and "base" for SH setting in the rest of this communication.

5. Result & Discussions

In the previous section we described how we designed and measured our feature set, specific to this task. Now we compute the within-group (Beg, Adv and CS) pairwise (Spearman's ρ) correlation between the seven features for each group for each of the settings [17]. We find a consistently high ($> \rho = 0.60$) correlation between ZCR and F0Percentage feature for all groups and all settings. As expected, it shows that there exists a strong correlation between the pitch based feature (F0Percentage) and the zero crossing rate (ZCR). On the other hand, we find a high correlation between word and V1 durations for the CS (Control Speakers) and Adv (Advanced learners) groups for both settings, but it does not follow for the beginner learner group for any setting. This shows that advanced learners master the lengthening of the vowel with respect to the fricative duration as French native speakers do. This is indeed a feature exploited to render voicing even if the percentage of non zero F0 frames is fairly low. Duration features are potentially important to discriminate among groups. Other than these two observations we do not remark any consistent pattern of high and significant correlations of the features taken in pairs. In this paper, we report all the values at 95% significance test level.

Since the histograms of the extracted features reveal that we cannot assume strict normality for all these cases, we perform the non-parametric Kruskal-Wallis rank sum test to select significantly different distributions [18]. We used the distributions of the three groups for each feature, for each setting in this test. These significantly different features (found through Kruskal-Wallis test) are marked in the Fig. 1 in red bracketed-stars. We notice that there are five significant features (namely, F0Percentage, Word-Dur(W), Fric-Dur(C), C/W and C/V1) for SR setting, and there are two significant features (namely, W and C) for SH setting. On the basis of the feature-wise boxplots, in the Fig. 1, between groups of each setting, we hypothesize that these significant features (which are mostly duration based features) would be useful to distinguish between the groups, whereas the pitch based feature ZCR, being always highly correlated with F0Percentage feature, may be a distinctive feature for voiced-unvoiced decision for each utterance [19, 20].

Agreement	SR	SH
Control Speaker	0.750	0.750
Beg	0.750	0.750
Adv	0.750	0.625

Table 1: Agreements between Expert and Non-Expert Ratings of three groups (Control Speaker, Beg & Adv) for two settings (SR & SH).

Before conducting feature based analysis we analyze the two kinds of voiced-unvoiced decisions from experts and non-experts for each speaker through the perception tests.

5.1. Expert Voiced & Unvoiced Perception Analysis

Two labelers experts in phonetics, who were unaware of our experiment, categorized the fricative realizations of learners and native speakers using two categories: voiced (+) and unvoiced (-). It shows very good inter-annotator reliability ($\sim 90\%$)[21]. From the annotators ratings for both settings we derive an overall team score for each speaker group. This is done through penalizing the mistakes with a negative scoring technique and normalizing it with the total number of cases. Therefore the team-score is computed as,

$$\text{Overall Team Score} = \frac{\text{count(voiced)} - \text{count(unvoiced)}}{\text{total count of participants}} \quad (1)$$

The goal to compute such kind of score is to obtain a score between +1 and -1. Also, this technique of scoring with penalization of negative performance, reflects the overall performance better for a small amount of samples [22].

We notice in the overall score displayed in Fig. 2a that there is a drop in performance of the L1 French native control speaker group at the repetition task, in comparison to the performance of read-only setting; here the half of the realizations is perceived as unvoiced and exhibits voicing measures in accord with this rating. Although the beginner group (Beg) was unable to perform well for SR Read-only setting, the performance improves during the SH repetition setting. But the performance of the Advanced learner group (Adv) remains the same for both settings. The experts have used Praat and did not focus only on the audio perception of the stimuli [16]. They actually used the spectrogram together with the display of the F0 curve. Furthermore, the manner of listening to the stimuli has a determining impact. In a phonetic annotation task, annotators use the gating paradigm which consists of extending a temporal window step by step and listening to it until it gives rise to the identification of the sound. This strategy presents the advantage of precise focusing on one sound independently of its context and at the same time decreases the perceptive fusion with the acoustic cues of

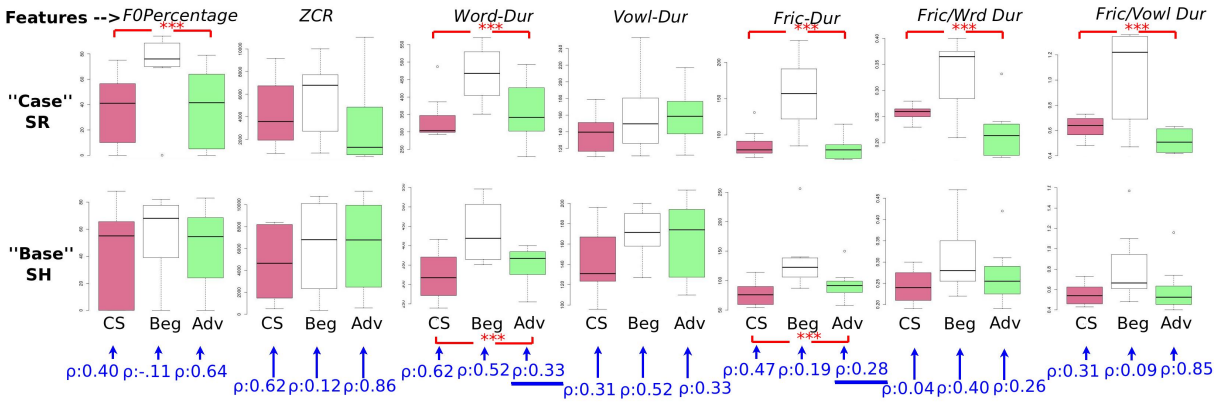
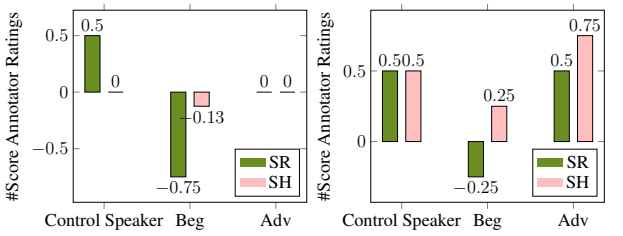


Figure 1: Comparison for boxplot of seven individual features of three groups (CS: Control Speakers, Beg, Adv speakers) for two settings; the red bracketed-stars denote the significant features from the Kruskal-Wallis test; the blue arrows with ρ values show the Spearman's correlation of a single feature for a single group, same colored boxplot distributions, across the both settings. The range of scale of features for both settings: F0Percentage(0,80); ZCR(2000,10000); Word-Dur(250,550); Vowel1-Dur(100,240); Fric-Dur(50,250); Fric/Wrd Dur(0.20,0.45); Fric/Vwl1 Dur(0.4,1.6).



(a) Score: Expert Ratings. (b) Score: Non-Expert Ratings.

Figure 2: Comparison of Overall Team-Score for Expert and Non-Expert Ratings for Three Groups (Control Speaker, Beg, Adv) for both (SR & SH) settings.

the previous vowel, especially its duration. This probably explains the rather low score of native speakers. So, we conduct a purely perceptive experiment with non-expert listeners.

5.2. Non-Expert Voiced & Unvoiced Perception Analysis

We collected data from 55 French native non-experts to whom we presented only the VCV forms of the words, to focus only on the production of obstruents (voiced stops and fricatives). Also we made the voiced fricative/stop rating choice more flexible with a five point Likert scale; additionally, we kept another question (in a five point scale) on the possible belonging to the L1 or L2 groups, for the confidence measurement. Each subject listened to 90 stimuli from the 1025 stimuli corresponding to a voiced obstruent in final position uttered by German learners of French and native French speakers in an equal number approximately. Unlike the experts, neither they accessed the spectrogram, nor they were informed of the feature values of the speech signal. We use majority voting to prepare the binary (voiced/unvoiced) decisions for all the non-experts [23]. The inter-annotator agreement is excellent (98%). We compute the overall team-score using the same Eqn. 1 using the non-expert ratings, the figure is shown as the Fig. 2b.

Agreement between Expert & Non-Expert Decision: We compute the agreement between the expert and non-expert ratings that is the average observed agreement across all annotators and speaker-subjects [24]. In most of the cases of Table. 1 there are satisfactory agreements between expert and non-expert ratings, except the advanced (Adv) learners in SH setting which presents a slight drop. In the next section, we analyze the results with respect to the feature-based decisions.

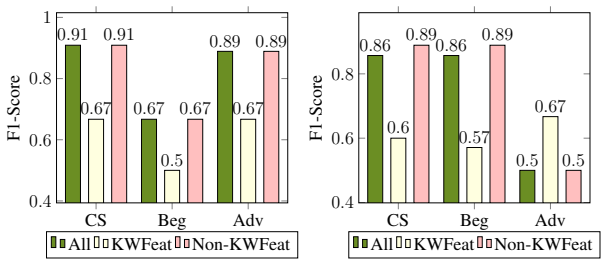
5.3. Comparison of Expert & Non-Expert Decision using Feature based K-means Clustering

In Sec. 5 the distinction between the three groups (Beg, Adv and CS) was discussed. Now our objective is to distinguish between voiced and unvoiced classes of realizations. Since our proposed feature set is devised to reflect the voiced-unvoiced property of acoustic realizations, we use the feature vectors to differentiate among the speakers into two classes, using K-means clustering, for each group of two settings [25]. Then we compute the relevance $F1$-score of the clustering results, with respect to the expert and non-expert ratings. The $F1$-score is the harmonic mean of precision and recall values, where its best value is 1 and worst is 0. Intuitively, the precision is the ability not to classify a negative label as positive, and the recall is the ability to classify all the positive labels [26]. The result is shown in the Fig. 3. There are three categories of result: 1. $F1$-score using all the seven features (All) 2. $F1$-score using significant features of Kruskal-Wallis test (KWFeat) and 3. $F1$-score using non-significant features of Kruskal-Wallis test (Non-KWFeat). Like [19, 20], we also found that ZCR and vowel features are effective for voiced-unvoiced distinction.

Among all the $F1$-scores, only the Adv group of SH setting shows lower $F1$ relevance. For the same case Table. 1 states worse performance for agreement. Maybe the Adv. learners exaggerated the production in the repetition task; in spite of this, we notice that the Adv learners slightly improved the voiced fricative production proficiency. In the Fig. 1 the Spearman's ρs (pointing boxplot distributions with blue arrows) show improvement for fricative and word duration for advanced learners between reading and repetition tasks (underlined values in Fig. 1). Ideally, ρs need to be less than the respective value of control speaker group. The ρs of the beginners show prominent improvements in productions, as an effect of the repetition task.

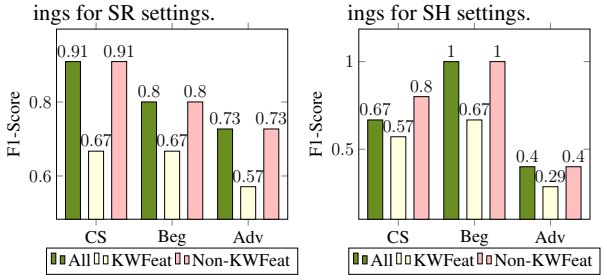
6. Feedback Correction

As shown above, the perception of the voicing phonetic feature mainly relies on the realization of voiced frames (i.e. positive F0 values) during the fricative, and a long vowel. Additional cues, as a fairly low intensity of the frication noise w.r.t. that



(a) Clustering w.r.t. Expert Ratings for SR settings.

(b) Clustering w.r.t. Expert Ratings for SH settings.



(c) Clustering w.r.t. Non-Expert Ratings for SR settings.

(d) Clustering w.r.t. Non-Expert Ratings for SH settings.

Figure 3: Comparison of F1-Score for Expert and Non-Expert Ratings versus Clustering decisions for Three Groups (Control Speaker:CS, Beginners:Beg, Advanced:Adv) for both (SR & SH) settings, with All, Kruskal-Wallis Significant (KWFeat) and Non-Kruskal-Wallis Significant Feature Set.

of the previous vowel and the presence of voicing in high frequency of the noise also contribute to the perception of voicing. The detection voicing during the fricative via F0 determination and the measurement of segment durations via the automatic segmentation provided by forced alignment are fairly robust and thus lead to a relevant phonetic diagnosis. The boundary between a vowel and a fricative is easily detected by ASR (Automatic Speech Recognition), and also more robustly since the expected error corresponds to an unvoiced fricative which is spectrally very different from the previous vowel. 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[14].

We describe this elaborated strategy of voicing correction. Adding voicing directly to the unvoiced fricative is possible by adding a synthetic voiced signal in low frequency. However, preliminary attempts showed us that this raises difficulties in generating a continuous phase at the boundary between the vowel and the fricative and degrades voice quality as well. We thus resort to TD-PSOLA [27] for concatenating a voiced fricative pronounced by a native speaker at the end of the learner's vowel and change the vowel duration as well. The voiced fricative is produced by a native/control speaker (from now on called teacher) in the same word, and therefore in the same phonetic context. It is a favorable situation because the concatenation intervenes between a vowel and a fricative, which exhibit very different spectral properties and also because both sounds are produced in the same vocalic context. Fig. 4 describes the implementation of this feedback. It is to be noted that the concatenation strategy that exploits pitch marks calculated on the learner's and teacher's speech signals, prevents the apparition of spectral discontinuities. The F0 curve of the teacher's signal is modified to ensure F0 continuity.

Two sets of teacher signals were recorded, one for each gender. This concatenation strategy turned out to be very efficient

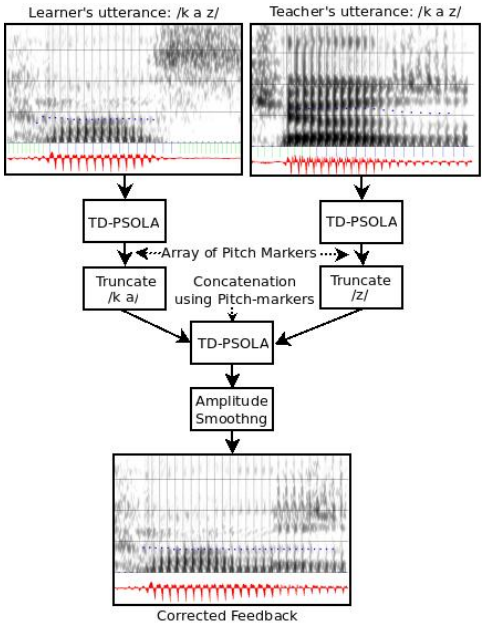


Figure 4: Process flow diagram for feedback correction with real spectrograms of /k a z/ with unvoiced /z/ (learner's) and of /k a z/ after revoicing of /z/. The original F0 curves The red dotted lines superimposed on the spectrogram represent the original and final F0 curves.

and perception tests targeting the intelligibility and voice quality shown that the concatenation sounds very natural.

7. 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. These results also confirm that the perception of the voicing feature is complex and incorporates several acoustic correlates. We found that: 1. the advanced learners are able to pronounce voiced fricatives better than the beginners because they are mastering the temporal scenario better (i.e. ratio of duration between vowel and consonant and ratio of voiced frames in the fricative segment). 2. repetition setting improves the quality of fricative pronunciation, most prominently for the beginner learners.

From an individual VCV realization point of view voicing (given by ZCR) and vowel duration enable the distinction between voiced and unvoiced realizations as judged by native speaker listeners. Since ZCR and duration features emerge as significant features to make this distinction, we used pitch-based TD-PSOLA method to design auditory feedback for the learners. In order 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. The advantage is to delimit the acoustic cues involved, and thus to improve the perceptive impact of the acoustic feedback proposed to learners.

8. Acknowledgement

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, 2013-2016.

9. References

[1] J. E. Flege, "Second Language Speech Learning: Theory, Findings and Problems ," pp. 233–272, 1995.

[2] J. E. Flege and R. Davidian, "Transfer and developmental processes in adult foreign language production," *Applied Psycholinguistic Research*, vol. 5, pp. 323–347, 1984.

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

[4] J. Sieczkowska, B. Möbius, and G. Dogil, "Specification in context - Devoicing processes in Polish, French, American English and German sonorants," in *Proceedings of Interspeech 2010 (Makuhari, Chiba, Japan)*, 2010, pp. 1549–1552.

[5] D. Jouvet *et al.*, "Analysis of phone confusion matrices in a manually annotated French-German learner corpus," in *Workshop on Speech and Language Technology in Education*, 2015.

[6] R. Wiese, *The Phonology of German*. Oxford: Clarendon Press, 1996.

[7] W. J. Barry, "The relevance of phonetics for pronunciation training. ," in *PHONUS 2, Working Papers Phonetics, Institute of Phonetics, University of Saarland*, 1996, pp. 5–19.

[8] O. Engwall, "Feedback from Real and Virtual Language Teachers," in *Working Papers 52*, 2006, pp. 41–44.

[9] K. Hirose, F. Gendrin, and N. Minematsu, "A pronunciation training system for Japanese lexical accents with corrective feedback in the learners own voice," in *EUROSPEECH*, 2003.

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

[11] ETSGlobal, "The Common European Framework of Reference for Languages," http://www.etsglobal.org/Fr/Eng/Research/CEFR.

[12] J. Trouvain, B. A., 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 (To appear)," in *10th Language Resources and Evaluation Conference (LREC)*, 2016.

[13] Y. Laprie, D. Jouvet *et al.*, "Project-Team PAROLE: Analysis, perception and recognition of speech," http://raweb.inria.fr/rapportsactivite/RA2011/parole/parole.pdf, INRIA-LORIA, Tech. Rep., 2015.

[14] 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.

[15] D. Fohr and O. Mella, "CoALT: A software for comparing automatic labeling tools ," in *LREC.2012*, 2012.

[16] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*, 5th ed., http://www.praat.org/, Jun. 2013.

[17] L. Breiman and J. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the Americal Statistical Association*, vol. 80, pp. 580–619, 1985.

[18] M. Hollander and D. A. Wolfe, *Nonparametric Statistical Methods*. John Wiley & Sons, 1973.

[19] J. Saunders, "Real-time discrimination of broadcast speech/music," in *proceedings of ICASSP 1996*. IEEE, 1996.

[20] G. S. Ying, C. D. Mitchell, and L. H. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement." Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on. Vol. 2. IEEE, 1993." vol. 2, 1993.

[21] O. Mella, D. Fohr, and A. Bonneau, "Inter annotator agreement for a speech corpus pronounced by French and German language learners," in *Workshop on Speech and Language Technology in Education SLaTE 2015*, 2015, pp. 143–148.

[22] P. McCoubrie, "Improving the fairness of multiple-choice questions: a literature review," *Medical teacher*, vol. 26, no. 8, pp. 709–712, 2004.

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

[24] R. Artstein and M. Poesio, "Inter-coder agreement for computational linguistics," *Computational Linguistics*, vol. 34, no. 4, pp. 555–596, Dec. 2008.

[25] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Applied Statistics*, vol. 28, pp. 100–108, 1979.

[26] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

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