IDIAP RESEARCH REPORT

PROSODY IN SWISS FRENCH ACCENTS: INVESTIGATION USING ANALYSIS BY SYNTHESIS

Pierre-Edouard Honnet [a] Alexandros Lazaridis [a]
Jean-Philippe Goldman Philip N. Garner

Idiap-RR-04-2014

MARCH 2014

[a] Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Prosody in Swiss French Accents: Investigation using Analysis by Synthesis

Pierre-Edouard Honnet[1], *Alexandros Lazaridis*[1], *Jean-Philippe Goldman*[2], *Philip N. Garner*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]University of Geneva, Geneva, Switzerland
{pierre-edouard.honnet, alaza, phil.garner}@idiap.ch, jean-philippe.goldman@unige.ch

Abstract

It is very common for a language to have different dialects or accents. The different pronunciations of the same words is one of the reasons for the different accents, in the same language. Swiss French accents have similar pronunciation to standard French, but noticeable differences in prosody. In this paper we investigate the use of standard French synthetic acoustic parameters combined with Swiss French prosody in order to evaluate the importance of prosody in modelling Swiss French accents. We use speech synthesis techniques to produce standard French pronunciation with Swiss French duration and intonation. Subjective evaluation to rate the degree of Swiss accent was conducted and showed that prosody modification alone reduces perceived difference between original Swiss accented speech and standard French coupled with original duration and intonation by 29%.

Index Terms: French accents, Swiss prosody, duration, intonation, speech synthesis

1. Introduction

The perception of different regional accents in a language can result from several factors. In French, the accents can vary because of different factors according to the regions. Eventhough there are noticeable differences at the pronunciation level of some phones between "Français de Référence" (FR) defined by Morin [1] as standard and Canadian French (or Quebec French) [2], between FR and Swiss French the differences in prosody have an important role in accent discrimination.

In automatic speech recognition (ASR), regional or foreign accents and dialects of a language bring variations that decrease performance of systems. It was shown by Huang [3] that the two main sources of inter-speaker variation were gender and accent.

Kat [4] proposes two solutions to overcome the variability introduced by accent: using a wide training database which includes accented data, or building accent-specific systems which will be used according to the accent of the speech to be recognised. In the literature, there are many other attempts to tackle the accent issue (mainly for non-native accented speech) in ASR by using adaptation techniques [5, 6, 7]. More generally, ASR systems are often confronted with non-native accents, and need to counteract effects of accent component.

Conversely, in text-to-speech synthesis (TTS), producing accented speech is desirable for some applications like speech-to-speech translation (S2ST), foreign language learning and dialect synthesis. Synthesising accented speech is still a quite new and challenging area. In most cases, different accents are modelled separately using different training data. There is only limited recent work on regional accent adaptation in TTS. Astrinaki [8] proposed interpolation of TTS models using closest speakers to a chosen geographical position. In this way, the English voice has average characteristics of these speakers representing the specific regional accent. Another work by Gutierrez [9] consists of generating intermediary accent transformations between native and foreign speakers, to evaluate pronunciation of learners (computer assisted pronunciation training). This research topic can be seen as part of TTS for under-resourced languages and cross-lingual speaker adaptation for TTS.

In the SIWIS [1] project, we are aiming at personalised S2ST, i.e. being able to recognise, translate to another language and synthesise speech with an output voice sounding like the input speaker's. One of the goals is to improve prosody rendering in the TTS part of the system. The final goal is a system enabling communication between speakers of different languages. As people are generally more comfortable when speaking to someone with the same accent as theirs, synthesising speech with the same accent is more convenient for the user.

Considering the differences in prosody between French and Swiss French speakers, we believe that using robust acoustic French TTS models combined with Swiss French prosody models will allow us to synthesise Swiss accented speech. Adapting acoustic level features from a language to an accent can be done using standard speaker adaptation techniques [10, 11]. On the other hand, adapting prosody is a much more challenging issue.

In this paper, a preliminary step towards this idea is to investigate whether it is feasible to produce speech with Swiss accent using standard French acoustic models. In this direction an attempt is made to explore the importance of prosody in Swiss French accent by evaluating the degree of accent of partially synthetic speech. Our hypothesis is that using only Swiss prosody modification allows to identify Swiss accents, even with standard French pronunciation. For that we combine standard French synthetic speech and Swiss duration and intonation. Native French listeners evaluated the accent of the resulting speech.

The rest of the paper is organised as follows. We first give an overview of the differences in prosody between FR and Swiss French. Then, we describe how to exploit French TTS models to synthesise Swiss accented French. A description of the data we used is made. Experiments are detailled in the following section. Finally we conclude and propose future directions.

2. Swiss French accent

It is important to underline that the differences between FR and Swiss French are limited, since the speakers are geographically close and furthermore, linguistically, Romandie (the French

[1]Spoken Interaction with Interpretation in Switzerland, http://www.idiap.ch/project/siwis/front-page

speaking part of Switzerland) would not be distinguished from Eastern and Southeastern France according to Knecht [12].

Consequently, in this paper, the focus is only on the acoustic aspects and prosody of Swiss accent and not on lexical differences nor on the grammatical or semantic structure. It is difficult to define a description of Swiss accent in a global way, mainly because there are different granularities of accent distinction within Switzerland. Most of the French native Swiss people can distinguish the accents by canton (administrative Swiss regions). Moreover, within a canton, people are even able to distinguish accents among cities. We attempt to give a quick overview of the shared peculiarities in Swiss accents.

Some differences in pronunciation between Swiss and French speakers exist, but according to Swiss regions, these are not equally strong. Metral [13] gives more details on segmental aspects of Swiss accents.

There are some divergences – as often in the area of prosody – on the rhythm topic, i.e. Swiss speakers are known to speak slower than French. Miller [14] showed that on read speech samples, speaking rate was the same for French and Swiss (from Vaud canton) speakers, but the articulatory rate (excluding pauses) was slower for Swiss speakers. French speakers use more pauses, which decreases their speaking rate. Schwab [15] recently lead an empirical study to verify whether Swiss people indeed speak slower than French people or not. The findings showed that pause frequency and duration were not different among some French, Belgian and Swiss speakers. However, articulation rate was found to be slower for Swiss speakers.

Schwab [16] compared two Swiss regional accents with French accent, regarding penultimate accentuation, shows that Swiss speakers are more likely to accentuate penultimate syllables than French speakers. Variations were also observed among Swiss regions with different strategies in expressing prominence on these syllables.

Swiss speakers are often said to produce more variations in their intonation, however it is hard to study the phenomenon due to the variety of intonation patterns. By accentuating different syllables, they generate different intonation patterns that may sound more "lively" to French listeners.

3. Combining standard French average spectral parameters with Swiss French prosodic features

Our hypothesis is that Swiss prosody plays a major role in Swiss accent discrimination and hence that adding prosodic information will help perceiving the Swiss accent. If this is true, the degree of accent should get closer to the original Swiss accented speech when Swiss prosody is added to non accented speech.

Using analysis by synthesis, we can modify some parts of a speech signal without altering other parts. Based on this principle, we propose to generate synthetic spectrum using French TTS models (corresponding to vocal tract in the source-filter model) and combine it with original duration and intonation of Swiss French.

3.1. Acoustic and prosodic features

The acoustic features commonly used in HMM-based speech synthesis were used: mel cepstral coefficients plus energy coefficient, pitch and band aperiodicity and the first and second derivatives for each feature. The duration information was estimated with forced alignment based on these features and the existing model parameters.

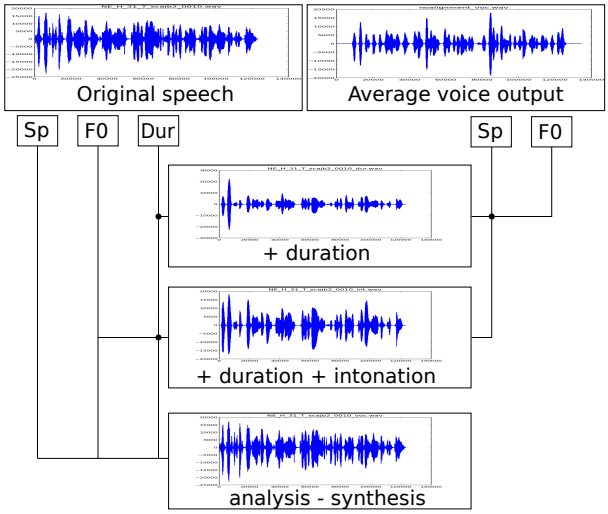

Figure 1: *Experimental setup. Features are Sp for spectrum, F0 for fundamental frequency, Dur for duration.*

3.2. Average French TTS models

We used statistical parametric speech synthesis models (hidden Markov models) to generate speech parameters that were used to "standardise" the pronunciation of all the speakers.

For that, we trained an average French HMM-based speech synthesis system using the HTS framework [17] and speaker adaptive training [18]. Average voice models are more robust than single speaker models and do not require extensive amounts of data from a single speaker.

Then, we could synthesise speech parameters and incorporate original duration and intonation information according to the prosodic features we wanted to evaluate using a vocoder.

3.3. Resynthesis

In parametric speech synthesis, the speech signal is represented by some parameters, spectrum and excitation in our case, and needs to be reconstructed to produce a waveform.

3.3.1. Vocoding

To homogenise the quality of the samples we want to produce, analysis synthesis was performed on the original waveforms, which consists of feature extraction and reconstructing the signal using a vocoder.

A completely synthetic waveform was also produced, using the French average models presented earlier. The parameters generated were finally put through the same vocoder.

3.3.2. Use of duration information

We used duration as a first prosodic feature to be added to the average synthetic speech. For that, we first extracted the duration information from the original waveforms using forced alignment: given the speech features and the corresponding transcription (full-context phonetic labels in our case), the Viterbi algorithm is used to estimate phone and state boundaries.

Using the state duration, a forced-aligned synthesis was performed, i.e. parameter generation given the known state sequence. The resulting speech was composed of synthetic parameters, but aligned in time with the original speech. The result was synthetic speech with original phoneme level duration.

3.3.3. Use of duration and intonation information

In this case, we also performed time alignment, and we replaced the synthetic intonation (log F_0) with the original one. After vocoding, the output was a speech signal composed of synthetic spectrum and aperiodicity coupled with original duration and intonation.

The reason for using both original intonation *and* duration is that it is not possible to use only original intonation, because the other parameters (spectral information) have to be aligned with the excitation part to reconstruct the speech signal.

Figure 1 gives an overview of the experimental setup described in this section.

4. Databases

The data used in this work comes from two databases: the BREF database [19] and a part of the PFC database [20].

BREF is a French read-speech corpus designed for speech recognition model training and testing. The sentences to be read by the speakers were chosen from the French newspaper *Le Monde*. It consists of recordings from 120 selected speakers (55 males and 65 females), recorded in a sound-proof room. The complete database represents more than 100 hours of speech. The speech from 10 male speakers was used for this work.

The dataset taken from the PFC database consists of read speech by Swiss French speakers and French speakers from Paris. These data have been recorded in 5 cities: Paris (France) and 4 cities in 4 different Swiss cantons, i.e. Martigny (Valais), Nyon (Vaud), Neuchâtel (Neuchâtel), and Geneva (Geneva). For each location, 4 male and 4 female speakers born and raised in the city were recorded. In this work, 12 speakers were selected among the 20 male speakers available.

5. Experiments

The experiment conducted consisted of the generation of partially synthetic speech combined with Swiss French prosodic information, and was evaluated with a subjective listening test.

5.1. French TTS model

The TTS models were trained on a subset of the BREF database composed of 6857 sentences (about 12 hours of speech) from 10 male speakers. We used 39 mel cepstral coefficient with energy coefficient, log F_0, 21 band aperiodicities extracted every 5 milliseconds with STRAIGHT [21] and their first and second derivatives. 5 state left-to-right HSMMs were trained with full-context labels using version HTS 2.1 and speaker adaptive training [18].

5.2. Subjective evaluation

One common sentence was selected for the 10 Swiss and 2 French male speakers from our PFC dataset. Only male speakers were used to match with our existing TTS average French male models. This sentence was used in previous Swiss accent related studies evaluations [22, 23].

> *"La côte escarpée du mont St Pierre connaît des barrages chaque fois que les opposants de tous les bords manifestent leur colère."*

It was segmented manually and the orthographic transcriptions were corrected manually before full-context label creation (adding pauses and hesitations). Features were extracted from Swiss French data the same way as for training data. The trained TTS models were then used to estimate the duration of Swiss speech data.

A listening test was conducted in order to evaluate the degree of accent of the file generated as described in section 3.3. For this purpose, a webpage was built enabling subjects to listen to 1 completely synthetic file and 1 file with original duration, 1 file with original duration and intonation and 1 vocoded file for each of the 12 speakers, which sums up to 37 files in total. The vocoded version is perceptually very close to the original recorded speech as it is only analysis synthesis. It enables us to allow for the vocoder effect in the perception. For each file, the listeners had to give a degree of Swiss accent between 1 and 5, 1 being "no accent" and 5 "strong accent" (in the instructions, "no accent" was defined as *standard accent* and close to Paris accent). The test took approximately 10 minutes, and the listeners could listen to the files as many times as they wanted.

28 subjects did the test. Among them, there were 17 males and 11 females, 23 were French and 5 were Swiss (2 from Vaud, 1 from Valais, one from Neuchâtel and one from St Gallen). Their age was distributed in four age ranges (8 were 19-25 years old, 10 were 26-35, 7 were 36-55 and 3 were 56 - 75).

5.3. Results

5.3.1. Degree of accent

Figure 2 shows the mean and standard deviation of the three versions of the file for each speaker – the fourth version displayed in black, which is identical for each speaker, corresponds to the average voice output. The means and variances show that when adding intonation and duration the values get closer to the vocoded version than just adding duration, and modifying only duration gives closer values than the average voice output, as we expected. For the speakers with highest degree of accent (based on the vocoded version), $NE75$, $NY31$, $NY32$, $NY59$ and $NY70$ ($PA86$ has different behaviour), the means of the *intonation + duration* version is still much lower than the vocoded one. $PA86$ is a 86 year old Parisian and although he does not have a Swiss accent, his accent was perceived as strong. The average voice being based on French accent and pronunciation, he has the same pronunciation as the average voice. Adding the prosody resulted in a degree of accent close to the original, explained by both correct prosody and pronunciation.

These results are confirmed by a Wilcoxon signed rank test which was perform for each speaker among the four versions presented (3, plus the baseline average voice), as the data is ordinal [24]. In the case of average version against vocoded version, 9 out of the 12 speakers have significantly different scores; $GE24$, $GE27$ and $PA33$, corresponding to the least accented speakers, are not significantly different. In the case of the version with duration information against the vocoded version, 7 still have significantly different scores: $MA24$ and $NE31$ are not significantly different. Finally, when adding original intonation, the 5 speakers mentioned before as *very different* from the vocoded version are significantly different. $GE55$ and $PA86$ are not significantly different for these versions.

5.3.2. Prosody effect on accent perception

Table 1 shows the means of absolute differences between scores per speaker. For each speaker, a comparison was made between 2 versions of the file among the average voice output (ave), the version including duration (dur), the version including duration and intonation (int) and the vocoded version which is the reference (voc). In 8 cases out of 12, the combination of duration and intonation is closer to the vocoded version (values in bold).

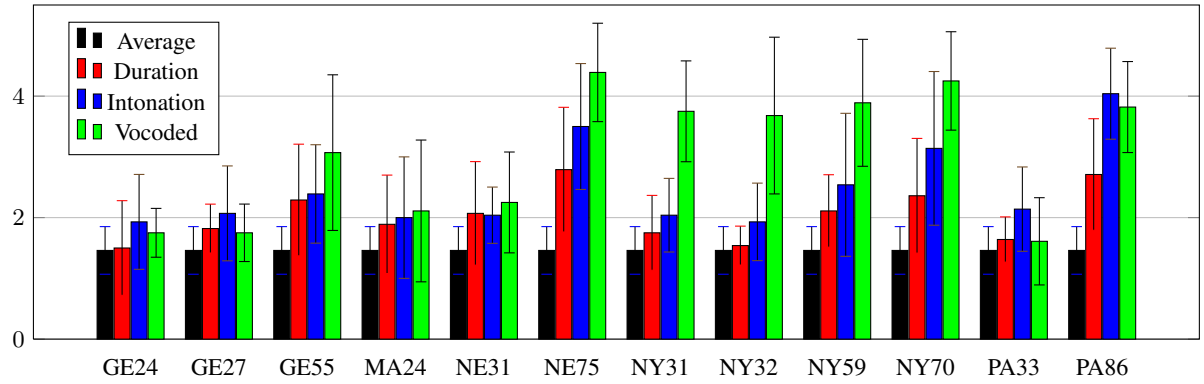

Figure 2: *Mean degree of accent for each version for the 12 speakers - Output of average TTS, with duration information, with duration and intonation, vocoded version*

Table 1: *Mean differences between configurations per speaker*

Speaker	$GE24$	$GE27$	$GE55$	$MA24$	$NE31$	$NE75$	$NY31$	$NY32$	$NY59$	$NY70$	$PA33$	$PA86$
ave_dur	1.04	0.82	0.96	0.96	0.96	0.86	0.89	0.96	0.93	0.79	1.29	0.64
ave_int	1.29	1.32	1.25	1.18	1.21	1.04	1.29	1.14	1.61	1.39	1.50	1.29
dur_int	0.61	0.79	1.07	1.14	0.75	0.68	0.61	0.96	0.89	1.11	1.00	0.93
ave_voc	1.68	1.93	1.54	1.25	1.57	1.54	1.75	1.89	1.60	1.60	2.21	1.79
dur_voc	**0.93**	1.54	**1.36**	1.64	**0.96**	1.04	**1.29**	1.57	1.39	1.39	1.64	1.57
int_voc	0.96	**1.25**	1.57	**1.07**	1.00	**1.00**	1.39	**1.46**	**1.21**	**1.14**	**1.21**	**1.21**

Table 2: *Mean differences between configurations*

	average	duration	intonation	vocoded
average	0	0.93	1.29	1.70
duration	X	0	0.88	1.36
intonation	X	X	0	1.21
vocoded	X	X	X	0

The 4 other cases give the advantage to the version including only duration information.

Table 2 gives the global absolute difference between each system. The last column gives the distance between the vocoded speech and the other versions. We can see that between the average voice output and the version with duration information we reduce the distance to the vocoded version by 20%, between the version with duration and the version including duration and intonation, the reduction is 11% and the overall improvement from average to duration and intonation version gives 29% improvement. A Wilcoxon signed rank test confirmed that the differences between score absolute differences were significant ($p - value < 0.01$ in the 3 cases).

It demonstrates that prosody plays an important role in Swiss accent perception. However, for the most accented speakers, prosody alone is not enough to obtain the same degree of accent. In these cases, adequate pronunciation is required to perceive the Swiss accent. This is backed up by the fact that accented Parisian speech can be produced with standard French pronunciation and specific prosody.

The low number of Swiss subjects did not allow us to evaluate the difference in accent perception between French and Swiss listeners, but the numbers showed similar trends for both groups.

6. Conclusions

In this paper we investigated the use of standard French pronunciation with Swiss prosody. This preliminary work was done with a view to adaptating French speech synthesis to Swiss accents. We hypothesised that Swiss accent was mainly characterised by its prosodic aspects. Analysis synthesis method and HMM-based speech synthesis were used to produce synthetic average French speech parameters which were then combined with natural speech prosodic features.

A subjective evaluation was conducted through a listening test to determine whether the degree of Swiss accent can be approached by modifying only the prosody of synthetic speech. The results showed that for 7 male speakers out of 12, using original duration and intonation with synthetic spectral parameters was not distinguished significantly from the original speech by the listeners. The difference of the scores between original speech and unaccented synthetic speech was significantly reduced by 20% by adding original duration and by 29% when adding original duration and intonation. This showed that prosody is important in the perception of Swiss accent. We also found that in the case of strong accents, prosody is not enough to model Swiss accent with standard French pronunciation. We did not use intensity information in this experiment, which would probably give further improvement.

Our future work will be to investigate the use of adaptation techniques for the pronunciation of the synthesis system to Swiss accents, and evaluate the impact of prosody in accent perception with Swiss pronunciation. The intonation contour could be investigated at a finer level to understand differences between regional accents. It would also be interesting to evaluate the accent identification rather than the degree of accent eventhough it is a difficult task for listeners.

7. Acknowledgements

This research is funded by the Swiss National Science Foundation under the SIWIS project – FNS Grant CRSII2_141903.

8. References

[1] Y. C. Morin, "Le français de référence et les normes de prononciation," *Cahiers de l'Institut de linguistique de Louvain*, vol. 26,

no. 1, pp. 91–135, 2000.

[2] M.-H. Côté, "Laurentian French (québec): extra vowels, missing schwas and surprising liaison consonants," in *Phonological variation in French: illustrations from three continents.*, R. Gess, C. Lyche, and T. Meisenburg, Eds. Amsterdam: John Benjamins, 2012.

[3] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," in *Proceedings of Eurospeech*, Aalborg, Denmark, 2001, pp. 1377–1380.

[4] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1. IEEE, 1999, pp. 221–224.

[5] S. Aalburg and H. Hoege, "Foreign-accented speaker-independent speech recognition." in *Proceedings of the International Conference on Spoken Language Processing*, 2004, pp. 1465–1468.

[6] X. He and Y. Zhao, "Fast model selection based speaker adaptation for nonnative speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 298–307, 2003.

[7] W. K. Liu and P. N. Fung, "MLLR-based accent model adaptation without accented data," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 3, 2000, pp. 738–741.

[8] M. Astrinaki, J. Yamagishi, S. King, N. d'Alessandro, and T. Dutoit, "Reactive accent interpolation through an interactive map application," in *Proceedings of the 8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, August 2013, p. 265.

[9] R. Gutierrez-Osuna and D. Felps, "Foreign accent conversion through voice morphing," Department of Computer Science and Engineering, Texas A&M University, Tech. Rep., 2010.

[10] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Speaker adaptation for HMM-based speech synthesis system using MLLR," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[11] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[12] P. Knecht, "Le français en Suisse romande: aspects linguistiques et sociolinguistiques." in *Le français hors de France*. Paris: Valdman, A., 1979, pp. 249–258.

[13] J.-P. Métral, "Le vocalisme du français en Suisse romande. considérations phonologiques," *Cahiers Ferdinand de Saussure*, no. 31, pp. 145–176, 1977.

[14] J. Sertling Miller, "Swiss French prosody: intonation, rate, and speaking style in the Vaud canton," Ph.D. dissertation, Graduate College of the University of Illinois, Urbana-Champaign, 2007.

[15] S. Schwab and I. Racine, "Le débit lent des suisses romands: mythe ou réalité?" *Journal of French Language Studies*, pp. 281–295, 2013.

[16] S. Schwab, M. Avanzi, J.-P. Goldman, P. Montchaud, I. Racine *et al.*, "An acoustic study of penultimate accentuation in three varieties of French," in *Proceedings of Speech Prosody*, 2012.

[17] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of the 6th ISCA Speech Synthesis Workshop*, 2007, pp. 294–299.

[18] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Transactions on Information and Systems*, vol. 90, no. 2, pp. 533–543, 2007.

[19] L. F. Lamel, J.-L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for french," in *Proceedings of EUROSPEECH*, 1991, pp. 505–508.

[20] J. Durand, B. Laks, and C. Lyche, *Phonologie, variation et accents du français*. Paris, Hermès, 2009.

[21] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[22] I. Racine, S. Schwab, and S. Detey, "Accent(s) suisse(s) ou standard(s) suisse(s) ? Approche perceptive dans quatre régions de Suisse romande." in *La perception des accents du français hors de France.*, A. Falkert, Ed., 2013, pp. 41–59.

[23] M. Avanzi, G. Christodoulides, S. S., B. A., and G. J.-Ph., "La variation prosodique régionale et stylistique en français – analyse de neuf points d'enquête PFC," in *Journées PFC*, Paris, 2013.

[24] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the blizzard challenge 2007 listening test results," in *Proceedings of Blizzard Challenge Workshop*, 2007.