Importance of Prosody in Swiss French Accent for Speech Synthesis

Pierre-Edouard Honnet, Philip N. Garner
Idiap Research Institute
<pierre-edouard.honnet, phil.garner@idiap.ch>

Abstract

En français, l'accent standard et l'accent suisse sont proches. Les proncia-tions observées sont effectivement similaires à quelques exceptions près. En revanche, la prosodie suisse a un effet important sur la façon dont les accents sont perçus. Nous étudions l'effet de la prosodie sur la perception de l'accent suisse sur de la parole synthétique par des locuteurs français natifs. Nous utilisons des modèles paramétriques de synthèse vocale français que nous adaptons à différents accents suisses. La prosodie de la synthèse est par la suite modifiée pour correspondre à celle de locuteurs réels. Nous montrons que le degré d'accent suisse est mieux reconnu en intégrant la prosodie adéquate.

Keywords: *Swiss French accents, text to speech synthesis, prosody.*

1. Introduction

1.1. Synthesis of regional accent

We are interested in synthesising regional Swiss accents in the context of speech to speech translation (S2ST) within Switzerland. As a first step in this direction, we investigate synthesis of Swiss accented French, in particular the perception of these accents regarding prosody.

In text to speech (TTS) synthesis, only limited recent work has been done on accent adaptation: Astrinaki et al. (2013) used many data from different locations within the United Kingdom and interpolated TTS models according to a target location. Gutierrez-Osuna and Felps (2010) generated accent transformations between native and foreign speakers, in order to evaluate pronunciation of learners.

1.2. Swiss prosody

The difference in perception among French accents can be due to several factors. If we compare standard French (Morin, 2000) with Swiss French, we need first to underline that differences are limited because of the geographic proximity, and they are linguistically close (Knecht, 1979). The pronunciation of Swiss speakers is slightly different from French speakers, but prosody also plays an important role in accent perception.

Métral (1977) describes the segmental aspects of Swiss accents and underlines that the differences observed between Swiss and French pronunciations are not equally strong among different Swiss regions.

As for the rhythm issue, Swiss speakers are known – outside of the scientific world – to speak slower than French speakers. The literature is somehow divergent on this aspect of prosody in Swiss and French speech, but Schwab et al. (2012) and Miller (2007) both showed that articulatory rate (excluding pauses) was significantly slower for Swiss speakers than for French speakers. Miller (2007) found that French speakers use more pauses, which brings their overall speaking rate closer to Swiss French speakers.

Schwab and Racine (2013) investigated accentuation, and showed that Swiss speakers are more likely to accentuate penultimate syllables than French speakers. These syllables were also found to be expressed in different ways among different Swiss regions.

At the intonation level, Swiss speakers are often known for having more variation in their intonation, but it is difficult to estimate with the intonation patterns being different. As they accentuate different syllables in different ways, their intonation sounds more lively to French listeners.

1.3. Swiss accent perception in synthesis

As prosody is important in Swiss French accent, we believe that to synthesise Swiss accented speech the prosody must be well modelled. In this direction, our previous work investigated the perception of the degree of accent of standard French pronunciation supplanted with Swiss prosody (Honnet et al., 2014). Our findings were that prosody helps to perceive the accent closer to the original speakers, but that for strong accent it was not enough.

Based on our previous results, we investigate how accent is perceived when the pronunciation is adapted to the accent of one of the French speaking region of Switzerland, and we provide the prosody of particular speakers from this region in the synthesis process. Our hypothesis is that adding prosody modification to adapted synthetic speech will improve accent perception.

In the rest of the paper, we first present how we integrate prosody in our accented speech synthesis, then results are presented and the last section concludes the paper.

2. Adapting TTS to regional accents and evaluating prosody

2.1. *Adaptation of the TTS models*

The speech synthesis method we use in this work is HMM-based speech synthesis. Based on our previous work (Honnet et al., 2014), we attempt to synthesise Swiss French with an adequate prosody. We previously used standard French intonation with Swiss prosody. In this work, we want to adapt the TTS models to target regional accents, in order to provide the pronunciation observed in these regions. The average French models are the same as in our previous work.

Based on our set of 12 speakers from 5 different locations (4 Swiss and 1 French), we adapt the French models to the accent using the data from all the speakers for each location and it results in average regional accent models. This is done with common speaker adaptation techniques (Yamagishi et al., 2009). For each speaker, we use 20 sentences for adaptation (leaving 1 sentence for evaluation). According to the location, we had different numbers of speakers to perform the adaptation. Paris had 2 speakers, Geneva 3 speakers, Martigny 1 speaker (in this case we actually do standard speaker adaptation), Neuchâtel 2 speakers and Nyon 4 speakers.

2.2. *Combining accented synthesis with target speakers' prosody*

Following the procedure described in our previous work, we use the original prosody from Swiss speakers (and French for comparison). We restrict ourselves to duration and intonation in this work. Using the same speakers as in 2.1., we created a test sentence in four different ways:

1. *Adapted:* it corresponds to the output of the adapted voice from the same location.

2. *Adapted + duration:* we use time information from original speech and other parameters from the adapted voice (same location).

3. *Adapted + duration + intonation:* duration information and intonation are taken from original speech, other parameters from the adapted voice.

4. *Vocoded:* it corresponds to the original speech passed through a

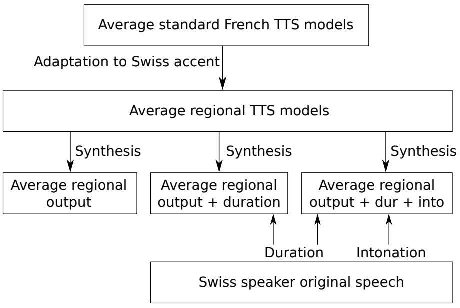
Average standard French TTS models
Adaptation to Swiss accent
Average regional TTS models
Synthesis
Synthesis
Synthesis
Average regional output
Average regional output + duration
Average regional output + dur + into
Duration
Intonation
Swiss speaker original speech

Figure 1: *Experimental setup.*

vocoder to achieve the same speech quality as other files.

By comparing these 4 different versions of the file for each speaker, we can evaluate the difference between using the actual prosodic cues from original speech and prosodic cues automatically generated from adapted model synthetic speech. The prosodic parameters are actually also adapted when adapting the average models to regional accent, but using several speakers' data for adaptation has a smoothing effect (to add to the smoothing effect of the parameterisation).

Figure 1 summarises the experiment described in 2.1. and 2.2..

The details about the models (features, data, etc.) and the tools we use can be found in our previous work.

2.3. *Subjective degree of accent evaluation*

To evaluate the accent of the synthetic and partly synthetic files produced, we conducted a webpage-based subjective evaluation, where the listeners had to rate the degree of accent between 1 (not accented) and 5 (strongly accented). The subjects had to listen to 36 files corresponding to version 2 to 4 of the test sentence described in 2.2., 5 files corresponding to version 1 (1 per location), and 1 file from standard French average TTS models, summing up to 42 files. The test took approximately 10 minutes.

19 French native listeners participated to the study. There were 7 Swiss (mainly from Valais and Vaud) and 12 French, 4 females and 15 males.

3. Results

We first look at the degree of accent of each speaker with respect to the conditions (adapted output, with duration, with duration and intonation, vocoded). Figure 2 gives the mean over the listeners of the degree

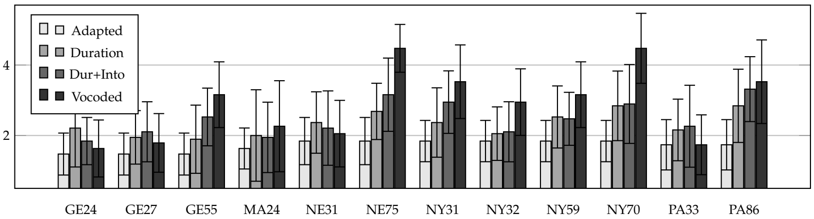

Figure 2: *Mean opinion score of the degree of accent for each version for the 12 speakers – Adapted TTS output, with duration information, with duration and intonation, vocoded version.*

of accent for each speaker in each configuration (for comparison, the average output with no adaptation score is 1.42).

A Wilcoxon signed rank test was performed for each speaker among the four versions. We found that for every speaker, there was no significant difference between the scores obtained for the vocoded version and for synthesis with duration and intonation information, no significant difference between the vocoded version and for synthesis with duration information, and no significant difference between the version with duration and the version with duration and intonation. In almost all the cases, there was a significant difference between each of these three versions and the adapted output with no prosodic information. We cannot conclude that there is no difference between our adapted models supplanted with original prosody, however we see that perceptually, adding prosodic cues improves the degree of accent of the synthetic speech and it cannot be distinguished significantly from original speech.

If we measure the absolute differences between scores per speaker, we observe that from average standard French output to regional accent adapted output with original duration and intonation, the mean distance is reduced by 41%, compared to the 29% obtained in our previous work. If we only use duration, the reduction is of 30%.

4. Conclusions

The results showed that there is no significant difference between the accent perceived for the original speech and the accent for adapted synthetic speech including original prosody (duration and intonation). We also got a reduction of 41% of the difference between the accent perceived in the case of an average model output and our system with original prosody. The hypothesis that we made is demonstrated as listeners could not distinguish significantly original speech from TTS adapted to regional accent provided with correct prosody in terms of degree of accent.

Acknowledgements

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS).

References

Astrinaki, M., Yamagishi, J., King, S., d'Alessandro, N., and Dutoit, T. (2013). Reactive accent interpolation through an interactive map application. In *Proceedings of the 8th ISCA Speech Synthesis Workshop*, page 265, Barcelona, Spain.

Gutierrez-Osuna, R. and Felps, D. (2010). Foreign accent conversion through voice morphing. Technical report, Department of Computer Science and Engineering, Texas A&M University.

Honnet, P.-E., Lazaridis, A., Goldman, J.-P., and Garner, P. N. (2014). Prosody in Swiss French accents: Investigation using analysis by synthesis. In *Proceedings of the 7th Speech Prosody Conference*, Dublin, Ireland.

Knecht, P. (1979). Le français en Suisse romande: aspects linguistiques et sociolinguistiques. In *Le français hors de France*, pages 249–258. Valdman, A., Paris.

Métral, J.-P. (1977). Le vocalisme du français en Suisse romande. considérations phonologiques. *Cahiers Ferdinand de Saussure*, (31):145–176.

Miller, J. S. (2007). *Swiss French prosody: intonation, rate, and speaking style in the Vaud canton*. PhD thesis, Graduate College of the University of Illinois, Urbana-Champaign.

Morin, Y. C. (2000). Le français de référence et les normes de prononciation. *Cahiers de l'Institut de linguistique de Louvain*, 26(1):91–135.

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

Schwab, S. and Racine, I. (2013). Le débit lent des suisses romands: mythe ou réalité? *Journal of French Language Studies*, pages 281–295.

Yamagishi, J., Kobayashi, T., Nakano, Y., Ogata, K., and Isogai, J. (2009). Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):66–83.