Detection of Phone Boundaries for Non-Native Speech using French-German Models

Dominique Fohr, Odile Mella

Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France

Abstract

Within the framework of computer assisted foreign language learning for the French/German pair, we evaluate different HMM phone models for detecting accurate phone boundaries. The optimal parameters are determined by minimizing on the non-native speech corpus the number of phones whose boundaries are shifted by more than 20 ms compared to the manual boundaries. We observe that the best performance was obtained by combining a French native HMM model with an automatically selected German native HMM model.

Index Terms: computer assisted foreign language learning automatic speech alignment, HMM

1. Introduction

The success of future systems for computer assisted foreign language learning relies on providing the learner personalized diagnosis and relevant corrections of its pronunciations. In such systems, a non-native speaker utters a word or a sentence and receives immediate feedback. For that the uttered sentence must be automatically segmented and phonetically annotated with high accuracy because a segmentation fault may lead to erroneous feedback or correction. High accuracy means to obtain an automatic phonetic alignment system that provides accurate temporal boundaries while being tolerant of non-native pronunciation deviations of the learner [1]. The aim of our study is how to obtain accurate temporal boundaries in the case of a bilingual French/German corpus.

2. Corpus IFCASL

The IFCASL (Individualised Feedback in Computer-Assisted Spoken Language learning) corpus is a bilingual speech corpus for French and German language learners. It was designed in order to allow an in-depth analysis of both segmental and prosodic aspects of the non-native production of these languages by beginners and advanced learners [2]. Each speaker had to perform several tasks in both languages L1 and L2. Among these tasks, he had to read aloud a sentence (29 sentences) and to read aloud a sentence after hearing this sentence pronounced by a native speaker (31 sentences). The speakers are adults or teenagers, female or male, and, beginners (A2 or B1 level) or advanced learners (C1). The recordings can be classified into 4 sub-corpora (GF, GG, FF, and FG). Three of them are used in this study:

- GF sentences: French sentences produced by 40 native German speakers;
- GG sentences: German sentences produced by the same 40 German speakers;
- FF sentences: French sentences produced by 50 native French speakers.

All sentences were automatically segmented and phonetically annotated. Then a part of these sentences was manually checked at the levels of phones (labels and boundaries) and corrections were made if necessary.

The aim of our study is how to obtain an accurate automatic phonetic alignment of the non-native GF sentences using if appropriate the other sub-corpora.

As the IFCASL corpus was designed to contain specific speech phenomena of interest for the French/German pair some of the words appear in multiple sentences. Therefore, for our study we split the GF corpus into two parts, GF-train (880 sentences) and GF-test (923 sentences), which do not contain the same vocabulary. We also split in the same manner the FF sentences into FF-train and FF-test corpora.

3. Methodology

To obtain an accurate automatic phonetic alignment system based on Hidden Markov Models (HMM), we choose a two-step methodology. First we determine the phone sequence that best represents the learner's utterance. Second, we determine the phone boundaries with a forced alignment using the sequence of phones determined in the previous step. The goals of the two stages are different: minimizing the number of deletions, insertions and substitutions of phones for the first step and minimizing the boundary shifts for the second one. Therefore, the optimal parameters (HMM models and phonetic lexicon) could be different.

In this paper, we are only interested in the second step in the context of non-native speech. We want to determine the optimal parameters by assuming that the sequence of phones obtained by the first step is perfect. For that, we use the sequence of phones from the manual labeling. The optimal parameters are determined by minimizing on the GF-test corpus the number of phones whose boundaries are shifted by more than 20 ms compared to the manual boundaries. We use our software CoALT (Comparing Automatic Labeling Tool) [3] for computing the boundary shifts. We define the following seven sets of models.

- **Native models**

French native HMM models are trained on the French radio broadcast news corpus: ESTER2 [4].

- **Native+Adapt_Native_auto models**

The previous models are adapted with Maximum Likelihood Linear Regression method (MLLR) on the FF-train corpus using the sequence of phones obtained by the automatic alignment with the French native models.

- **Native+Adapt_Native_manu models**

The native models are adapted on the FF-train corpus using the sequence of phones coming from the manual labelling.

- **Native+Adapt_Non-Native_auto models**

The native models are adapted on the non-native GF-train corpus using the sequence of phones obtained by the automatic alignment with these native models. The sentences of the speaker which will be aligned are removed from the GF-train.

- **Native+Adapt_Non-Native_manu models**

The native models are adapted on the GF-train corpus using the sequence of phones coming from the manual labelling.

- **Parallel_auto models**

Every model is composed of two HMM models in parallel: a French *native+adapt_native_auto* model and a German model. German models are first trained on the native German corpus Kiel [5] and then adapted on all the GG sentences (using the automatic alignment), except those of the speaker which will be aligned. For every French model, the German model put in parallel is automatically chosen. Given the set of N_G German HMM models, for a French model M_F, we build N_G sets of models. Each set is composed of all the French models except M_F, and, a parallel model (M_F and a German model). We align the GF-train corpus with each of these N_G sets of models. Finally, among the N_G models tested, the German model which will be parallel with M_F is the one (if it exits) that best improves the alignment according to our criterion of minimizing the boundary shifts.

- **Non-native models**

Non-native models are trained on the GF-train corpus except the sentences of the speaker which will be aligned, and, using the manual labelling (phones + boundaries).

4. Results

The different models are evaluated on the non-native GF-test corpus totaling 29400 phones. The audio files are parameterized with MFCC (Mel Frequency Cepstral Coefficient) and a 10ms frame shift. The HMM acoustic models have three states except for stops for which we tested models with 1 or 2 states for the closure and for the burst. According to our criterion of minimizing the boundary shifts, two-state models for both closure and burst were better and we have kept these models for the following experiments.

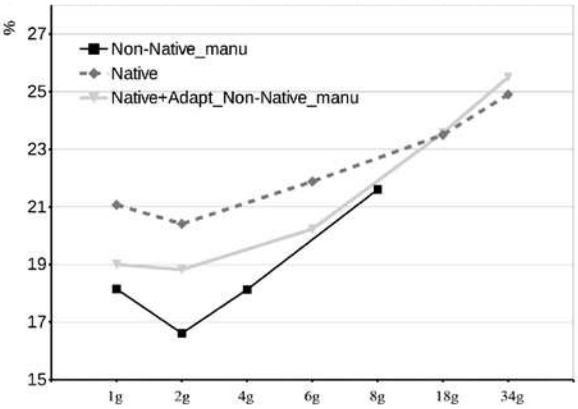

Figure 1: Percentage of shifts of boundaries > 20 ms according to the number of Gaussians per state.

We then determine the number of Gaussians that improves the accuracy of the boundaries. Figure 1 shows the percentage of shifts of boundaries greater than 20 ms for three sets of models. We can note that the shape of the curves are similar regardless the type of models and the optimal result is achieved for two Gaussians per state. This confirms the result that acoustic phone models with only a few Gaussian provide a better temporal precision than detailed acoustic models [6].

The number of Gaussian being set, we then evaluate seven types of models. The performance of each model is presented in Table 1. The confidence interval at the 95% confidence level is ± 0.5%. As expected, the French native models are the worst. These are the models that were originally used to automatically label the GF and FF corpora.

The best performance is obtained by the models trained on a non-native corpus but this requires a fairly big non-native speech corpus manually labeled at the phone level which is very costly. The best trade-off consists in putting in parallel two native models trained on native corpora and adapted on the environment of the computer assisted foreign language learning system. Moreover, we can see that adapting the models with manually-labeled data rather with automatically-labeled data does not improve significantly the accuracy of the boundaries.

Table 1: Percentage of shifts of boundaries > 20 ms.

Models	Shift >20ms
Non-Native_manu	16.6%
Parallel_auto	17.8%
Native+Adapt_Non-Native_manu	18.8%
Native+Adapt_Non-Native_auto	19.3%
Native+Adapt_Native_manu	19.4%
Native+Adapt_Native_auto	19.6%
Native	20.4%

5. Conclusion

In this study, we evaluated different HMM phone models for the second step of the alignment process: detecting accurate phone boundaries within the framework of computer assisted foreign language learning. The best performance was obtained by using phone models built by putting in parallel a French native HMM model and an automatically selected German native HMM model.

6. Acknowledgements

This work has been supported by an ANR/DFG Grant “IFCASL” to the Speech Group LORIA CNRS UMR 7503 – Nancy France and to the Phonetics Group, Saarland University –Saarbrücken Germany, 2013 –2016.

7. References

[1] S.M. Witt, “Automatic Error Detection in Pronunciation Training: Where we are and where we need to go,” *Proceedings of International Symposium on automatic detection on errors in pronunciation training,* vol.1, 2012.

[2] C. Fauth, and al. “Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process,” *LREC, Reykjavik, Iceland,* 2014.

[3] D. Fohr and O. Mella, “CoALT; A Software for Comparing Automatic Labelling Tools,” *LREC Istanbul, Turkey,* 2012.

[4] S Galliano, G Gravier, L Chaubard, “The ester 2 evaluation campaign for the rich transcription of French radio broadcasts.”, “ *INTERSPEECH* Brighton, UK, 2009.

[5] J. Kohler, “Labelled Data Bank of Spoken Standard German - The Kiel Corpus of Read/Spontaneous Speech”, *ICSLP*, *Philadelphia, USA*, 1996

[6] D. Toledano and L. Gomez, “Automatic Phonetic Segmentation”, *IEEE Trans. on Speech and Audio Processing,* v11, n6, pp. 617—625. 2003