Inter-annotator agreement for a speech corpus pronounced by French and German language learners

Odile Mella, Dominique Fohr, Anne Bonneau

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

This paper presents the results of an investigation of inter-annotator agreement for the non-native and native French part of the IFCASL corpus. This large bilingual speech corpus for French and German language learners was manually annotated by several annotators. This manual annotation is the starting point which will be used both to improve the automatic segmentation algorithms and derive diagnosis and feedback. The agreement is evaluated by comparing the manual alignments of seven annotators to the manual alignment of an expert, for 18 sentences. Whereas results for the presence of the devoicing diacritic show a certain degree of disagreement between the annotators and the expert, there is a very good consistency between annotators and the expert for temporal boundaries as well as insertions and deletions. We find a good overall agreement for boundaries between annotators and expert with a mean deviation of 7.6 ms and 93% of boundaries within 20 ms.

Index Terms: inter-agreement annotator, non-native speech alignment, computer assisted foreign language learning, German/French corpus, comparing labelling tool

1. Introduction

The success of future systems for computer assisted foreign language learning relies on providing the learner with personalized diagnosis and relevant corrections of its pronunciations. In such systems, the primary objective is to provide the learner with automated feedback which derives from an analysis of the learner's utterance and targets specifically the acoustic features to be improved [1]. For that purpose, the uttered sentence must be automatically segmented and phonetically annotated with high accuracy since a segmentation fault may lead to erroneous feedback or correction. High accuracy requires an automatic phonetic alignment system that provides accurate temporal boundaries while being tolerant of non-native pronunciation deviations of the learner.

Within the framework of the IFCASL[1] project [2], a large speech corpus of native and non-native speech for the French-German language pair was designed. This corpus is intended to developing and validating: (i) diagnosis and feedback algorithms, (ii) automatic phonetic alignment systems. For both tasks a precise segmentation is mandatory.

This corpus was automatically segmented and phonetically annotated by our speech-text alignment tool and then manually checked. As it is often the case for large corpus the verification has been done by several annotators. This final manual annotation is the starting point which will be used both to improve automatic segmentation algorithms and derive diagnosis and feedback. Therefore, it is necessary to know the degree of inter-annotator agreement. The aim of this paper is to evaluate this agreement.

Inter-annotator agreement is most often reported in terms of what percentage of the manual boundaries are within a given time threshold. With regard to native speech, Hosom in [3] reported results from different studies. These studies dealt with Italian, German, and English continuous speech, and, analyzed the consistency between few annotators (from two to four). He concluded that there is a fairly good agreement between human labelers across language and channel conditions with an average agreement of 93.8% within 20ms with a maximum of 96% for highly-trained specialists using rigorous and well-defined conventions. Concerning non-native speech, Gut and Bayer measured the reliability of manual annotations of speech corpora, made by six annotators, and have shown that manual annotation can be very reliable but depended upon the coding complexity [4].

2. Corpus IFCASL

2.1. Corpus description

The IFCASL 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. The corpus was recorded by fifty French learners of German and forty German learners of French in their native and second languages. The non-native speakers were classified by L1 teachers in three categories: beginners, intermediate and advanced. The corpus consists of four sets of sentences, corresponding to different speaking conditions: (1) reading sentences (about 30 sentences, referred to as SR sentences); (2) repeating sentences (about 30 sentences, referred to as SH), (3) focus elicitation, and, (4) reading of a short text. The two last parts of the corpus were not used in this study.

[1] Individualised Feedback in Computer-Assisted Spoken Language learning (IFCASL), supported by ANR (Agence Nationale de la Recherche and DFG (Deutsche Forschungsgemeinschaft)

2.2. Automatic and manual labeling of the corpus

All the SR and SH sentences were automatically segmented and phonetically annotated by our speech-text alignment tool based on acoustic Hidden Markov Models.

A part of the aligned sentences were manually checked at phones and words levels (phonetic transcription) and corrections were made if necessary. The French sentences uttered by German and French speakers were corrected by seven French annotators (undergraduate students in phonetics), managed by a French expert phonetician (an assistant professor in phonetics). Annotators must add or remove a phone label and change a label when the speaker uttered a speech segment different from the canonical pronunciation. In case of voicing or devoicing of a consonant, annotators must add a diacritical mark. Moreover, they must carefully verify the phone boundaries and move them if necessary. When the boundary set by an annotator would be arbitrary he/she should use a diacritical symbol to mark it as fuzzy (as, for instance, the boundary between /a/ and /R/ in the word "*départ*").

Since the phonetic segmentation has been checked and corrected by seven annotators, it was necessary to verify the consistency of the seven annotators with the expert annotator.

3. Inter-annotator agreement

3.1. Methodology

To verify the consistency of the seven annotators with the expert annotator, 18 audio files were selected and annotated by each of the seven annotators and by the expert phonetician. Among these 18 audio files, 12 were recorded by German learners (GF) and 6 by French speakers (FF). The audio files correspond to 13 different sentences (7 SH and 6 SR) with a total of about 625 phones. We used the software CoALT (Comparing Automatic Labelling Tool) to compare the results of the annotators to those of the expert annotator.

CoALT compares the results obtained by several labelers (automatic speech-text alignment tools or human labelers) with a reference alignment in order to rank them and display statistics about their differences. CoALT presents the advantage of allowing users to define their own comparison criteria [5].

The analysis of deviations made by non-native speakers often requires accurate temporal boundaries. In the case of German learners speaking French, for example, we paid special attention to: aspiration of voiceless stop consonants, final devoicing of obstruent consonants and vowel duration. In the same way, the analysis of rhythm and accents (lexical or focus) requires reliable boundaries. Therefore, we begin the inter-annotator agreement analysis in terms of shifts of boundaries.

3.2. Inter-annotator agreement regarding shifts of boundaries

For each sentence corrected by an annotator, CoALT first matches the sequence of phones with the sequence obtained by the expert annotator, using an elastic comparison algorithm that takes into account labels and time boundaries. Then, CoALT computes the boundary shifts between two matching phones if either both phones are identical or their substitution is allowed by a rule. The boundaries marked as fuzzy by the expert annotator have not been taking into account in this study. The expert annotator characterized 52 limits as fuzzy for a total of 625 labels. Finally, CoALT computes some statistics on the shifts.

3.2.1. Overall estimate of inter-annotator agreement

As a first overall estimate of the inter-annotator agreement, Table 1 shows for each annotator, the mean absolute shift of the boundaries computed on all the phones of the 18 sentences. It corresponds to about 520 boundaries per annotator. We can observe a fairly good overall agreement between the annotators and the expert. Therefore the annotation of the IFCASL corpus can be used to develop and assess new automatic segmentation tools. However, our results show that it will not be possible to require an automatic boundary accuracy better than ± 10 ms.

As the threshold of 20 ms is commonly used to compare the performance of human and automatic labelers, we computed the percentage of labels whose boundaries are shifted by less than 20 ms with respect to the boundaries set by the expert annotator. The average percentage of 93%, with a confidence interval at the 95% confidence level of ± 2.2%, corresponds well with the results reported by Hosom for native speech [3]. On the one hand, the agreement may have been slightly facilitated in some cases by the fact that annotators had started from the automatic alignment. However, the annotators were instructed to adjust any incorrect boundaries and place them as precisely as possible. But, on the other hand, the task was more complex because of the non-native speech.

Table 1. *Shifts of boundaries for each annotator.*

Annotator	Mean absolute shift (ms)	Shift <= 20ms
#1	7.1	94.1%
#2	9.1	90.1%
#3	7.6	93.3%
#4	6.7	95.3%
#5	8.0	93.2%
#6	7.4	91.9%
#7	7.6	92.9%
all	**7.6**	**93.0%**

3.2.2. Comparison between native and non-native speech

Figure 1 presents the percentage of labels whose boundaries are shifted by less than a threshold and compares them for native (FF) and non-native speakers (GF). As expected, we see that the segmentation of non-native speech is slightly more difficult than that of native speech. This is mainly due to the lack of fluency of most non-native speakers, which generates hesitations, insertions of speech segments such as glottal stops and fricatives /?, h/. The presence of erroneous realizations such as aspirated voiceless stop consonants (French voiceless stops are not aspirated) also explains this difference.

With CoALT, users can define classes for phones and their context in order to provide shift histograms. We used this feature to investigate the inter-annotator agreement in specific cases.

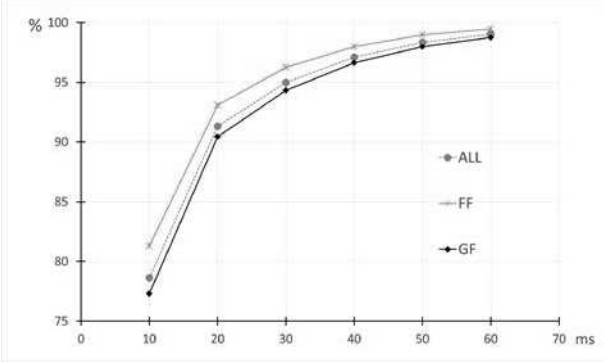

Figure 1. *For all labelers, percentage of labels whose boundaries are less than a time threshold from those of the expert labeler. The time threshold is indicated on the x axis.*

3.2.3. Shift of boundaries for stop consonants

One of the aims of the IFCASL project is to study French stop consonants pronounced by German learners. Therefore, the boundaries of the closure and the burst must be as reliable as possible in order to provide a relevant diagnosis and a good feedback to the learner. We suppose that there is a relationship between the agreement rate and the difficulty of the task. Figure 2 shows the histograms of shifts for voiced and unvoiced stops between (1) closure and burst and (2) burst and vowel, computed on all sentences. In each case the total number of occurrences is indicated in parentheses.

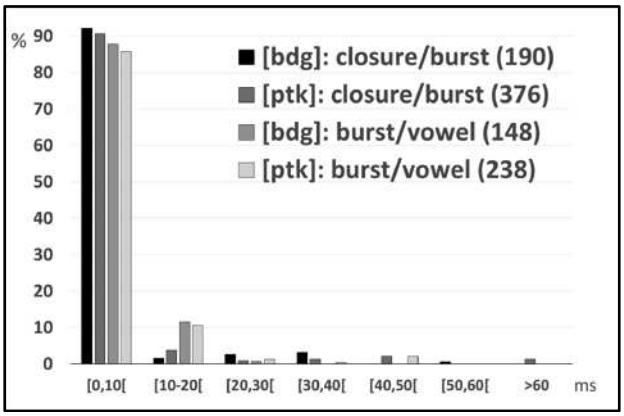

Figure 2. *Histogram of boundary shifts for stops.*

As expected, limits between closure and burst are easier to set than those between burst and vowel. We observe that most shifts are inferior to 10 ms (about 89% of the cases). Depending on the boundary category, the confidence interval at the 95% confidence level is between ± 3% and ±5%. This is a very satisfying result since a good temporal precision is necessary for French stop bursts, which are relatively short.

3.2.4. Shift of the vowel boundaries according to the context

Within the framework of language learning for the French-German pair, vowel duration analysis is important to evaluate lexical accent, vowel quantity (which exists in German but not in French) as well as fluency. Thus, reliable vowel boundaries are mandatory. Figures 3 and 4 show histograms of shifts of the limits between vowels and different classes of consonants.

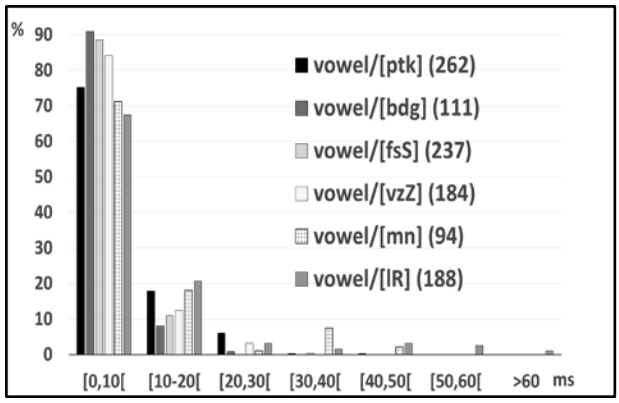

Figure 3. *Histogram of the shifts of the boundaries between vowels and their right context.*

Let us comment the results for which the boundaries shifts are the most important in both contexts: /m, n/ and /l, R/.

Results for /l,R/ were expected since these consonants have a very clear, vowel-like formant structure when they are in vocalic contexts. Such results could have been worse since CoALT excludes the boundaries considered as fuzzy by the expert annotator (but not those of the student annotators). On the other hand, results for /m,n/ were somewhat unexpected since the boundaries between nasals and vowels can be put with a good precision when the sentences are pronounced by French speakers. We believe that the relative lack of precision observed in this context is due to nasalization of vowels by German speakers. Indeed, French speakers, who have oral and nasal vowels in their phonological system, tend to preserve the phonemic contrast between them. In languages with no nasal vowels, such as English or German, oral vowels in contact with nasal consonants undergo a greater degree of nasal coarticulation [6]. The vowel nasalization leads to an unsteady vocalic signal difficult to segment, which may explain the divergence between the annotators. Hence, caution should be taken when using these specific boundaries.

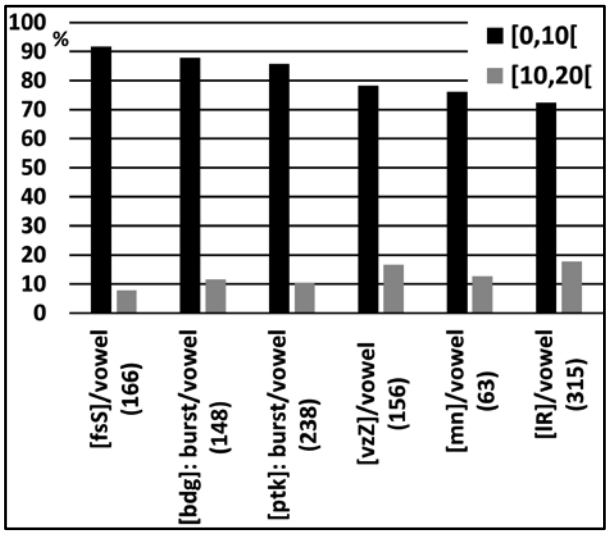

Figure 4. *Percentage of boundaries between vowels and their left context for [0, 10ms[and [10, 20ms[.*

3.3. Inter-annotator agreement regarding phone labels

3.3.1. Voicing/devoicing diacritics

There are two major differences between German and French systems with respect to the [voice] feature. The first difference between French and German is related to the phonetic implementation of the [voice] feature for stop consonants [7]. To be short, the presence vs absence of voicing due to vocal fold vibration is an important cue (not the only one) in the distinction of French /b,d,g/ vs. /p,t,k/, whereas the absence vs. presence of aspiration is an important cue for the same distinction in German. Voicing during closure is not mandatory for German /b,d,g/, whereas French /p,t,k/ are not aspirated. Hence, German speakers might realize the closure of French /b,d,g/ without glottal buzz, and /p,t,k/ with aspiration.

The second one is phonological and concerns final devoicing in German: In German, the opposition between voiced (/b,d,g,v,z,Z/) and voiceless (/p,t,k,f,s,S/) obstruents is neutralized in final position in favour of the realization of voiceless consonants [8], whereas in French the [voice] feature is distinctive in final position. This difference between both systems is known to be a source of error for German speakers, who tend to produce voiceless obstruents in final position when speaking French instead of the expected voiced consonants [9].

Both phenomena, the absence of (expected) periodicity during stop closure, and the absence of (expected) periodicity during the production of an obstruent in final position, have been indicated at the phonetic level by a "_0" diacritic added of the expected segment.

Note that the annotators have only checked periodicity (generated by vocal fold vibration) for both phenomena, and that the possible shift between categories due to final devoicing is not indicated (more than one cue is involved in such a shift).

Table 2 presents the agreement concerning the devoicing diacritic for voiced obstruent consonants. For every obstruent and every annotator we counted when the annotator agreed (or not) with the expert annotator about the absence or presence of the diacritic. The number of times the seven annotators and the expert one were in agreement is indicated in bold. Overall the percentage of agreement on the devoicing diacritic is good (88.5%). But for these 18 audio files, the addition of a devoicing diacritic by an annotator is correct only in 61% of cases. This result reflects the difficulty of the task particularly for non-expert annotators.

Table 2. *Agreement of the devoicing diacritic for voiced stops and fricatives between the seven annotators and the expert.*

		Annotators	
		without	with
Expert	without	**381**	46
	with	13	**71**

3.3.2. Insertions and deletions

Regarding phones labels, we have not analyzed the overall rate of substitutions because the annotators had instructions to focus their corrections on the phone boundaries and on the voicing and devoicing of obstruent consonants. With regard vowel timbre, confusions concern essentially mid-close and mid-open vowels, which are not easy to detect. Thus we ask annotators not to take too much time on this phenomenon.

We can observe in Table 3 that there is a very low rate of deletions and insertions between the seven annotators and the expert. Recall that the 18 sentences have a total of 625 phones.

40% of insertions or omissions concern the schwa. This result is rather expected because schwa is often a very short and weak vowel whose presence is sometimes difficult to detect.

Table 3. *Percentage of insertions and deletions for each annotator.*

Annotator	**Insertions**	**Deletions**
#1	1.4%	1.8%
#2	1.0%	1.4%
#3	1.8%	2.4%
#4	1.4%	1.4%
#5	1.4%	1.4%
#6	0.8%	1.6%
#7	0.8%	1.0%
all	**1.2%**	**1.6%**

4. Conclusions

Within the framework of the IFCASL project, a speech corpus of native and non-native speech for the language pair French-German was designed and recorded. Then, the automatic alignment of the audio files corresponding to the French and German speakers uttering French sentences (4100 audio files) were manually checked by a group of seven annotators. The corpus will be used for developing and assessing automatic algorithms that will provide the diagnosis of the learner mispronunciations (see first results on phone confusions in [10]) and the corresponding feedback. Therefore, in this paper, we analyzed the inter-annotator agreement according to an expert annotator for boundary shifts, insertions and deletions as well as devoicing diacritic. Whereas results for the presence of the devoicing diacritic show a certain degree of disagreement between the annotators and the expert, there is a very good consistency between annotators and the expert for temporal boundaries as well as insertions and deletions. Indeed, the mean absolute shift computed on all phones is less than 10 ms.

We can also conclude that the large IFCASL corpus with its manual labeling is well-suited for the development and the assessment of new automatic phonetic alignment systems for non-native speech and it is a good starting point for developing diagnosis and feedback algorithms.

5. 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.

6. 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, A. Bonneau, F. Zimmerer, J. Trouvain, B. Andreeva, V. Colotte, D. Fohr, D. Jouvet, J. Jügler, Y. Laprie, O. Mella, B. Möbius. "Designing a Bilingual Speech Corpus for French and German Language Learners: a Two-Step Process," *LREC 2014- Language Resources and Evaluation Conference, Reykjavik, Iceland, Proceedings,* 2014.

[3] J.P. Hosom, "Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information," *Ph.D. thesis*, *Oregon Graduate Institute,* May 2000.

[4] U. Gut and P.S. Bayerl "Measuring the Reliability of Manual Annotations of Speech Corpora", *Speech Prosody, Speech Prosody, Nara, Japan*, pp565-568.2004

[5] D. Fohr and O. Mella, "CoALT; A Software for Comparing Automatic Labelling Tools," *LREC 2012- Language Resources and Evaluation Conference, Istanbul, Turkey, Proceedings*, 2012.

[6] S.Y. Manuel, "The role of contrast in limiting vowel-to-vowel coarticulation in different languages". *The Journal of the Acoustical Society of America, 88*, 1286-1306. 1990.

[7] L. Lisker and A. Abramson, "A cross-language study of voicing in initial stops" Word, pp. 384-422, 1964.

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

[9] A. Bonneau, "Realizations of French voiced fricatives by German learners," *accepted in ICPhS Glasgow*, 2015.

[10] D. Jouvet, A. Bonneau, J. Trouvain, F. Zimmerer, Y. Laprie, B. Möbius "Analysis of phone confusion matrices in a manually annotated French-German learner corpus" *Submitted to this workshop, Slate*, 2015.