Reliability of non-native speech automatic segmentation for prosodic feedback

Larbi Mesbahi, Denis Jouvet, Anne Bonneau, Dominique Fohr, Irina Illina, Yves Laprie

Speech Group, INRIA - LORIA, 615 rue du Jardin Botanique, 54602 Villers les Nancy, France

{larbi.mesbahi,denis.jouvet,anne.bonneau,dominique.fohr,
irina.illina,yves.laprie}@loria.fr

Abstract

This paper investigates the reliability of phonetic boundaries obtained through automatic segmentation of non-native speech for automatic prosodic feedback for foreign language learning. Indeed, prosodic feedback requires checking the fundamental frequency and the duration of phonetic segments of the learner utterances with respect to some reference patterns. Segmentation evaluations carried out on non-native speech data show that the automatic segmentation process takes benefit from the introduction of non-native pronunciation variants, and that several phonetic boundaries obtained through automatic segmentation seems to be reliable enough for providing relevant prosodic feedback. This concerns for example boundaries between obstruent sounds, such as plosives and fricatives, and vowel sounds.

Index Terms: Language learning, automatic phonetic segmentation, non-native speech.

1. Introduction

In the last decade there has been enormous progress in the domain of assisted foreign language learning (e.g. [1], [2]). The interest comes from the diversification of resources and tools for learning, and also from social and economic motivations for development. However, it is recognized that the effectiveness of these means could be increased with better speech algorithms for detecting pronunciation deficiencies and for providing specific feedback with respect to these deficiencies [3]. One of the main problems in foreign language learning is the automatic localization and detection of the mispronunciations [4]. This requires speech recognition technologies that are robust to non-native speech. Methods have been proposed to derive goodness of pronunciation scores [5] based on likelihood metrics. Such systems take benefit from the introduction of acoustic models of the subject's native language, as well as expected mispronunciations. Although the non-native mispronunciations are dependent on the mother tongue, attempts have been made in developing mispronunciation detection approaches that are widely independent on the learner's mother tongue [6].

As the prosody is important for speech understanding, melodic curve visualization was used in second language learning and its impact tested on learners [7]. The Fluency project was aimed at providing feedback on duration errors [8]. Prosodic information was also used among the scoring features in [6]. Although some learning tutor systems provide prosodic feedback, they typically just play or replay the sound of a native speaker (teacher) uttering the same word or sentence as a reference. Winpitch LPL [9], a speech signal editor, proposes functions especially designed for L2 teachers and learners, which enable the user to modify by hand fundamental frequency and duration and annotate prosodic displays. An innovative approach was proposed in [10] with the goal of improving both perception and production. The main idea of the approach was to combine a detailed automatic prosodic feedback explaining what is incorrect and how to correct it with an auditory feedback. The auditory feedback is based on an automatic transformation of the learner utterance with prosodic parameters that match those of the native reference speaker. The transformations are based on an improved version of the TD-PSOLA method [11] and the Winsnoori software [12] is used for visualization, analysis and processing of the speech signals. Of course, the audio signal corresponding to the native reference speaker utterance is also available.

Prosodic feedback relies on the comparison of prosodic parameters computed on the learner utterance to some reference patterns, usually the same prosodic features estimated on some native reference speaker utterance. Prosodic features typically include the fundamental frequency and the duration of the sounds. However, to get the duration of the sounds a phonetic segmentation is necessary. Some boundaries are difficult to set, and even human experts may disagree in some cases, as for example for boundaries between liquids or semi-vowels and vowels. Automatic segmentation, through forced alignment of the speech signal with the sequences of acoustic models corresponding to the known or possible pronunciation variants, usually provides good results for native speech signals, although it is not always perfect. However, here, we have to deal with non-native speech phonetic segmentation, in which the acoustic variability is higher. Moreover, it is well known that when returning feedback in training tutors, the system must avoid wrong feedback, especially feedback that erroneously mention an error when the pronunciation is actually correct. Consequently, as the quality of the prosodic feedback depends strongly on the quality of the phonetic segmentation, this paper focuses on investigating the reliability of the phonetic boundaries achieved through automatic phonetic segmentation on non-native speech signals.

The paper is organized as follows. Section 2 describes the non-native speech data used, including the baseline phonetic segmentation process and extensions for handling non-native speech data. Section 3 deals with the alignment process used for comparing the automatic segmentation results with the reference phonetic segmentation. Besides a baseline alignment process based on the edit distance, an improved version is proposed that takes into account the temporal position of the automatic and reference phonetic segments. Finally, section 4 analyses the phonetic segmentation results, and focuses on the analysis of the phonetic boundary errors. A conclusion ends the paper.

2. Non-native speech data

2.1. Description of the non-native speech corpus

The corpus we used was originally devoted to a prosodic study (project INTONALE [13]). It is made up of 50 English sentences, (essentially assertions, but also questions) varying in length and complexity, as well as 11 other sentences without syntactic constraints that were used as "filler" sentences. Thirty four speakers recorded the corpus. This corresponds to 29 female and 5 male speakers, being about 20 years old, and studying French literature at the University of Nancy. The subjects were divided into four groups, each group reading a different part of the first corpus (the set of 50 sentences). The filler sentences were read by all the speakers. This gave a total of about 800 sentences.

All the sentences were recorded through an in-house recording platform. During the recording sessions, the intensity level was checked and the subject was invited to record again the sentence when the intensity level was either too low or too high. Each sentence of the corpus was displayed on the computer screen, written in bold characters, and was preceded and possibly followed by a short text giving a "situation". The subjects did not have at their disposal the pronunciation of the sentences by native English speakers (they only have the text displayed on the computer screen). They could repeat the sentence, listen to the recording, cancel it and record it again as often as they wanted. However, they seldom used that option.

Due to the time required for manually segmenting the non-native speech data, the utterances of only eighteen randomly selected speakers have been manually segmented and, consequently, used for evaluating the quality of the automatic phonetic segmentation. The manual segmentation was carried out using speech visualization tools (WinSnoori [14] and Wavesurfer [15]).

2.2. Baseline automatic segmentation

The baseline automatic segmentation is obtained through forced alignment of the speech utterance with the sequences of HMM (Hidden Markov Models) acoustic models corresponding to the possible pronunciation variants for that sentence.

For the baseline system, native pronunciation variants were used. They were obtained from the CMU pronunciation dictionary [16] that relies on a set of 48 phonemes. The acoustic models were trained using the TIMIT speech corpus [17], a native English high quality speech corpus collected from more than 600 speakers. As it is frequently observed that context-independent HMM provides better segmentation results than context-dependent HMM [18], context-independent acoustic models have been used.

Using these native acoustic models and the native pronunciation variants extracted from the CMU lexicon, leads to the "native"-based segmentation process.

2.3. Extensions for non-native speech segmentation

As we deal here with non-native speech uttered by foreign language learners, we have to take into account non-native mispronunciation in the automatic segmentation process. Non-native mispronunciations come from the fact that a person speaking in a foreign language may replace some sounds of the foreign language by sounds they find similar in their mother tongue, and also may omit or insert sounds. Moreover, some extra errors may occur because the learner does not know the correct pronunciation of some words.

Hence, non native variants have been introduced in the pronunciation lexicon. However, as, in a first step, our goal was focused on analyzing the quality of the phonetic boundaries, we did not spend time developing sets of generic rules for inferring the pronunciation variants, but directly defined by hand the pronunciation variants of the words according to the experience gained during manual segmentation of the data. The impact of ruled based approaches will be investigated in further studies.

Some pronunciation variants involve sounds of the mother tongue, in our case French sounds, such as /y/ and /ɑ̃/ which do not exist in English, but are used by learners reading the English text, especially when they do not know the "good" pronunciation. The acoustic models for those units were trained on a subset of the ESTER speech corpus [19]. The context-independent acoustic models of these French sounds were then used in addition to the context-independent acoustic models of the native English speech, such combination being frequent in speech recognition systems dealing with non-native speech, in addition to introducing non-native pronunciation variants [20], [21].

3. Aligning sequences of phonetic segments

Aligning sequences of phonetic segments is the process used for comparing the automatically derived phonetic segments with the manually defined phonetic segments. This alignment is necessary for evaluating the quality of the automatic phonetic segmentation.

3.1. Standard alignment of phonetic sequences

A standard solution for aligning two sequences of symbols resides on using the edit distance based dynamic programming. The process computes the minimum distance between the two sequences of symbols, which corresponds to the minimum number of insertions, deletions and substitutions that are necessary for transforming one sequence into the other.

The dynamic programming process involved, compares partial sequences of the automatic and reference sequences, and iteratively extends the length of the partial sequences that are compared. This relies on the following formula:

$$D(i,j) = min\begin{cases} D(i,j-1) + C_{ins}^{e} \\ D(i-1,j-1) + d^{e}(i,j) \\ D(i-1,j) + C_{del}^{e} \end{cases}$$

where $D(i,j)$ is the distance between the i first symbols (phonemes) of the reference and the j first symbols (phonemes) of the automatic segmentation; the edit cost $d^{e}(i,j)$ is equal to C_{sub}^{e} or C_{cor}^{e} depending on whether the phoneme i of the reference and the phoneme j of the automatic segmentation are different or equal; and C_{ins}^{e} and C_{del}^{e} are respectively the insertion and deletion costs.

The alignment leading to the minimal global cost provides the alignment between the automatic and the reference phonetic sequences. This method usually provides good alignment, especially when the number of mismatches is low. However, when aligning a non-native speech segmentation to a manual reference segmentation, the number of discrepancies can get high, and may lead to undesirable alignments as the one displayed in Figure 1, where red thick solid line arrows show the mis-aligned segments (whether insertion, confusion and deletion), and light dash arrows show correct alignments.

The alignment in Figure 1 corresponds to the utterance "Hopscotch amuses Maria". The actual pronunciation exhibits the mispronunciation of the learner ("amuse" instead of "amuses"). However, as, here, the automatic segmentation was conducted using the native pronunciation lexicon, the learner was expected to pronounce "amuses" correctly. The forced alignment has thus inserted the segments required by the lexicon pronunciation, even if they were not actually pronounced. Figure 1 also shows that the phonemes /z/ that are aligned are not the ones that matches the best from the temporal point of view. Such undesirable correspondences impact on the evaluation of the phonetic boundaries.

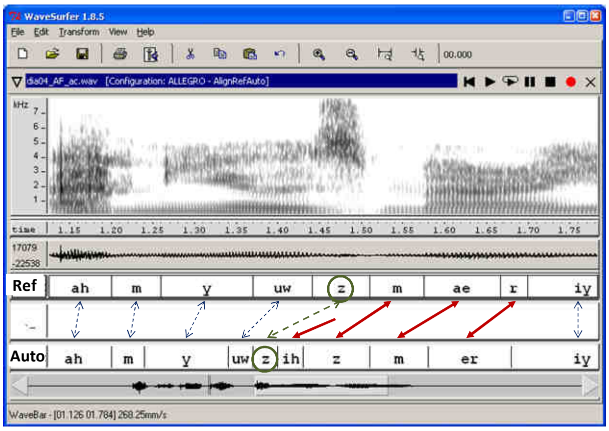
WaveSurfer 1.8.5
File Edit Transform View Help
dia04_AF_ac.wav [Configuration: ALLEGRO - AlignRefAuto]
Ref
ah
m
y
uw
z
m
ae
r
iy
Auto
ah
m
y
uw
z
ih
z
m
er
iy
WaveBar - [01.126 01.784] 268.25mm/s

Figure 1: *Example of undesirable segment alignment.*

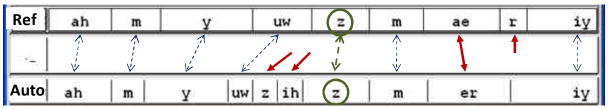

Figure 2: *Example of improved segment alignment thanks to temporal information.*

3.2. Temporal information in aligning segments

In order to avoid such undesirable alignments between automatic and reference phonetic segments, the alignment function was slightly modified to include a temporal penalty factor:

$$D(i,j) = min \begin{cases} D(i,j-1) + \lambda \cdot C^e_{ins} + (1-\lambda) \cdot C^t_{ins} \\ D(i-1,j-1) + \lambda \cdot d^e(i,j) + (1-\lambda) \cdot d^t(i,j) \\ D(i-1,j) + \lambda \cdot C^e_{del} + (1-\lambda) \cdot C^t_{del} \end{cases}$$

where $d^t(i,j)$ is a penalty score which gets larger when the phoneme i of the reference and the phoneme j of the automatic segmentation gets further apart in time. C^t_{ins} and C^t_{del} are respectively the associated insertion and deletion costs. λ is a weighting factor (chosen equal to 0.80) that specifies the compromise between the edit-based scores and the temporal-based scores.

Figure 2 displays the results of the improved alignment process on the previous example. The results shows a more natural correspondence between the /z/ segments of the automatic and reference phonetic sequences.

4. Phonetic segmentation quality

The analysis of the phonetic segmentation quality was conducted using phonetic classes corresponding to vowel sounds (Vow.), semi-vowels (Sem.), liquids (Liq.), plosives (Plo.), fricatives (Fri.), affricates (Aff.), nasal consonants (Nas.) and the other remaining miscellaneous units (Misc.).

4.1. Impact of non-native pronunciation variants

The impact of taking into account non-native pronunciation variants is showed through the confusion matrices displayed in Tables 1 and 2 below. The tables exhibit the number of matches inside classes (for diagonal elements) and between classes (off diagonal). The last line and column of each table gives the number of insertions (#INS) and deletions (#DEL) respectively. Bold figures in Table 2 corresponds to improved values.

Improvements are observed when non-native pronunciation variants are introduced in the pronunciation lexicon and thus used in the automatic segmentation process. In particular the total amount of insertions is largely reduced from 471 (with native variants only) to 335 when non-native variants are introduced, for just a small increase in the total number of deletions (from 371 to 422).

Table 1. *Class-based confusion matrix using only native pronunciation variants during segmentation.*

	Vow.	Sem.	Liq.	Plo.	Fri.	Aff.	Nas.	Misc.	#DEL
Vow.	4207	4	8	6	5	0	13	106	143
Sem.	3	229	5	3	0	0	0	6	20
Liq.	1	0	870	4	0	0	1	31	93
Plo.	2	0	0	1824	7	11	1	1	8
Fri.	3	2	3	4	1638	0	2	1	11
Aff.	0	0	0	1	3	85	0	0	0
Nas.	6	0	1	3	4	0	1222	0	11
Misc.	11	0	7	4	1	0	1	299	85
#INS	94	9	31	100	55	0	16	166	

Table 2. *Class-based confusion matrix using native and non-native pronunciation variants during segmentation.*

	Vow.	Sem.	Liq.	Plo.	Fri.	Aff.	Nas.	Misc.	#DEL
Vow.	**4324**	**1**	**4**	**4**	**1**	0	**6**	**10**	**142**
Sem.	3	227	5	**2**	0	0	0	**1**	28
Liq.	5	0	**896**	**3**	0	0	1	**1**	94
Plo.	**0**	0	1	1782	9	**10**	1	**0**	51
Fri.	4	2	**1**	6	1623	1	2	**0**	25
Aff.	0	0	0	1	**0**	**88**	0	0	0
Nas.	**3**	0	1	3	**2**	0	**1227**	0	11
Misc.	46	5	17	**1**	1	0	1	266	**71**
#INS	109	**3**	36	**61**	**35**	1	17	**73**	0

4.2. Reliability of phonetic boundaries

A detailed analysis of the phonetic boundary errors was then conducted. The automatic phoneme boundaries were compared to the reference phoneme manual boundaries. In order to get the best possible view of the boundary errors, the number of boundaries for which the errors fall in a given time interval were counted. Several time intervals were considered:]-∞,-35ms],]-35,-25ms],]-25,-15ms],]-15,-5ms],]-5,+5ms],]+5,+15ms],]+15,+25ms],]+25,+35ms] &]+35,+ ∞[. For several phonetic classes, the percentage of boundaries for which the temporal errors fall in each time interval are displayed for the beginning (Figure 3) and end (Figure 4) of plosive sounds, and for the beginning (Figure 5) and end (Figure 6) of fricative sounds. Several curves are displayed in each figure according to the class of the preceding or of the following sound. Only pairs of classes for which at least one hundred phonetic boundary occurrences are observed in the non-native speech corpus were kept and displayed in the Figures.

The figures show that, for the considered phonetic classes, a large amount of the boundary errors are limited to the interval [-15ms, +15ms], and that only a very small number of the considered phonetic boundaries falls outside the [-25ms, +25ms] interval. Consequently this shows that the automatic phonetic boundaries between obstruent sound (plosives and fricatives) and vowel sounds are quite reliable and such information could be used to provide relevant prosodic feedback for foreign language learning.

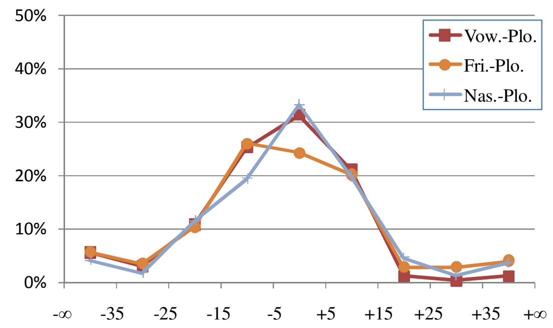

Figure 3: *Boundary errors for the beginning of a plosive according to the preceding class.*

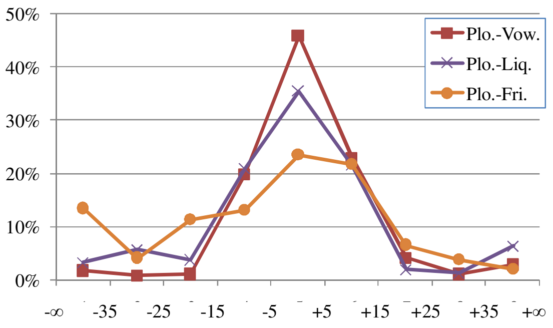

Figure 4: *Boundary errors for the end of a plosive according to the following class.*

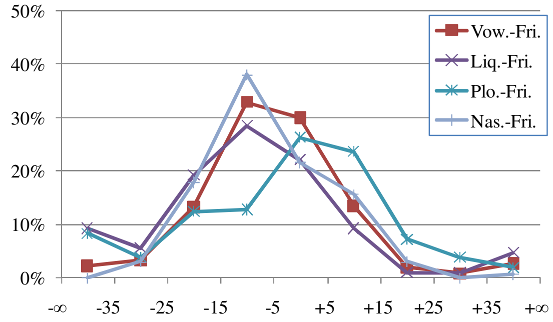

Figure 5: *Boundary errors for the beginning of a fricative according to the preceding class.*

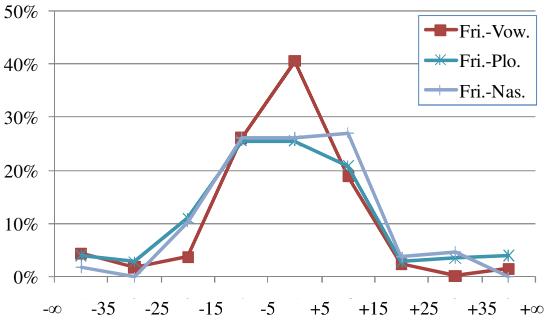

Figure 6: *Boundary errors for the end of a fricative according to the following class.*

5. Conclusions

This paper has investigated the quality of automatic phonetic boundaries on non-native speech, in view of their usage for prosodic feedback in foreign language learning.

Prosodic feedback requires comparing prosodic features of the learner's utterance with respect to some reference patterns. One set of prosodic features corresponds to the duration of the sounds. The estimation of the duration requires a precise phonetic segmentation of the learner's utterance. Experiments showed that when dealing with non-native speech, it is very important to include non-native pronunciation variants in the pronunciation lexicon. Future work will investigate further the usage of rules for deriving those variants, as well as the potential impact of introducing too many unnecessary variants.

A detailed analysis of the phonetic boundary errors showed that for some phonetic classes, the boundaries are quite reliable. This means that they could be used for estimating relevant phoneme or syllable durations, and thus for providing relevant prosodic feedback to the learners. However, as the reliability of the boundaries depends on the phonetic classes, care should be taken in defining the teaching exercises, and/or in the level of detail in the prosodic feedback.

6. References

[1] *Speech Communication*, special issue on "Spoken language technology for education", vol. 51, 2009.

[2] Eskenazi, M., "An overview of spoken language technology for education", *Speech Communication*, vol. 51, pp. 832-845, 2009.

[3] Stouten, F. & Martens, J.-P. "On the use of phonological features for pronunciation scoring", *Proc. ICASSP'2006*, Toulouse, France, vol. I, pp. 329–332, 2006.

[4] Herron, D., Metzel, W., Atwell, E., Bisiani, R., Daneluzzi, F., Mortan, R. & Schmidt, J., "Automatic localization and diagnosis of pronunciation errors for second language learners of english", *Proc. EUROSPEECH'1999*, Budapest, 1999.

[5] Witt, S. M. & Young, S.J., "Phone-level pronunciation scoring and assessment for interactive language learning", *Speech communication*, vol. 30, pp. 95–108, 2000.

[6] Hacker, C., Cincarek, T., Maier, A. Hebler, A. & Noth, E., "Boosting of prosodic and pronunciation features to detect mispronunciation of non-native children", *Proc. ICASSP'2007*, Honolulu, Hawai, USA, vol. IV, pp. 197–200, 2007.

[7] Vardanian, R., "Teaching English Intonation through Oscilloscope Displays", Language Learning, vol. 14, 1964.

[8] Eskenazi, M., Ke, Y., Albornoz, J. & Probst, K., "The fluency pronunciation", Proc. InSTIL'2000, Dundee, 2000.

[9] Martin, P., "WinPitch LTL II, a Multimodal Pronunciation Software". *Proc. InSTIL/ICALL*, Venice, Italy, 2004.

[10] Henry, G., Bonneau, A., Colotte, V., "Tools devoted to the acquisition of the prosody of a foreign language", *Proc. ICPhS'2007*, Saarbrücken, Germany, 2007.

[11] Colotte, V. & Laprie, Y., "Higher pitch marking precision for TD-PSOLA", *Proc. EUSIPCO'2002*, Toulouse, France, 2002.

[12] Laprie, Y., "SNOORI, a software for speech sciences", *Proc. MATISSE'1999*, 1999.

[13] http://mathilde.dargnat.free.fr/INTONALE/intonale-web.html

[14] http://www.loria.fr/~laprie/WinSnoori/

[15] http://www.speech.kth.se/wavesurfer/

[16] ftp://ftp.cs.cmu.edu/project/speech/dict/

[17] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N. & Zue, V., "TIMIT Acoustic-Phonetic Continuous Speech Corpus", LDC, Philadelphia, 1993.

[18] Toledano, D., Gomez, L. & Grande, L., "Automatic phonetic segmentation," IEEE Transaction on speech and audio processing, vol. 11, pp. 617–625, 2003.

[19] Galliano, S., Gravier, G., and Chaubard, L., "The ESTER 2 evaluation campaign for rich transcription of French broadcasts", *Proc. INTERSPEECH'2009,* Brighton, UK, 2009.

[20] Bouselmi, G., Fohr, D. & Illina, I., "Combined acoustic and pronunciation modelling for non-native speech recognition", *Proc. INTERSPEECH'2007*, Antwerp, Belgium, 2007.

[21] Bartkova, K., & Jouvet, D., "On using units trained on foreign data for improved multiple accent speech recognition", *Speech Communication*, vol. 49, n°. 10-11, pp. 836-846, 2007.