Detecting relevant acoustic events for piloting improvement of intelligibility

Vincent Colotte, Yves Laprie

DETECTING RELEVANT ACOUSTIC EVENTS FOR PILOTING IMPROVEMENT OF INTELLIGIBILITY

Vincent Colotte and Yves Laprie
LORIA, Campus scientifique, BP 239, F-54506 Vandœuvre-lès-Nancy, FRANCE
Tel: +33 (0)3 83892074; fax: +33 (0)3 83413079
e-mail: Vincent.Colotte@loria.fr , Yves.Laprie@loria.fr

ABSTRACT

This paper presents a speech signal transformation which slows down speech signals selectively and enhances some important acoustic cues. This transformation can be used for hearing aids and also for second language acquisition by facilitating oral comprehension. Selective slowing down relies on the use the TD-PSOLA synthesis method. The strategy used to control slowing down exploits a spectral variation function which locates rapid spectral changes. The enhancement simply consists of amplifying stop bursts and unvoiced fricatives. These acoustic cues are detected automatically through the examination of energy criteria. This approach was evaluated in the context of second language acquisition, more precisely by evaluating improvements in oral comprehension through the transformation, the detection and a perceptive experiment. The detection of relevant events gives good results and experiments show that the oral comprehension is improved.

1 INTRODUCTION

We are working with a view of improving speech intelligibility by slowing down speech selectively. Our approach rests on the TD-PSOLA method (Time Domain Pitch Synchronous Overlapp and Add). In a previous paper [5], we presented our pitch marking algorithm and the resynthesis method. This papers deals with the strategy accepted to pilot the selective slowing down.

We believe that slowing down should operate at the phonetic level (unlike [8] who operate at prosodic level) in the case of hearing aids as well as in the case of tools for oral comprehension of a foreign language. Indeed, the method proposed by A. Nakamura et al.[8] does not seem to be easily transposable for French. In French, the importance of information brought by the breath group does not follow the fundamental frequency declination as it seems to be the case in Japanese TV broadcasts. In addition, modifications at the phonetic level allow slowing down to focus on specific aspects : only one kind of sound difficult to be perceived can be transformed. The strategy for applying slowing down must give rise to an automatic algorithm and should be piloted by the detection of some acoustic cues.

Two directions are then possible. Either the signal is segmented into phones, or the signal is marked at instants[1] of high concentration of acoustic cues, i.e. regions where speech acoustic features vary strongly and quickly.
Segmentation into phones can be implicitly achieved during speech recognition or can result from a manual phonetic transcription : both solutions were dismissed, the former for computation heaviness and accurateness, and the latter, although possible for oral comprehension exercises in language learning, because we want to implement a completely automatic method.
The second approach relies on the localisation of regions with a high concentration of acoustic cues. Unlike Liu [6] who uses energy criteria and expert knowledge on the features of events to be located, we use a method which assesses acoustic variations of speech. This method, called *Spectral Variation Function*, (proposed by G. Flammia and al.[2] and F. Brugnara and al.[1]) use mel-cepstre analysis. A coefficient, reflecting the spectral variation rate, is calculated for each window (of 20 ms every 10 ms) by considering neighbouring windows (L frames at left and at right). The function is given by the following formula :

$$SVF(t) = \frac{1}{2}(1 - \frac{1}{L^2}\sum_{i,j}\cos(V_i, V_j))$$

where $t - L \leq i < t$ and $t < j \leq t + L$ with t the referent instant and V_i is the vector which its components are the first six mel-cepstrum coeficients obtained on the frame i (V_i is also normalized with respect to the $2L+1$ frames of the calculation). The multiplicative constant $\frac{1}{L^2}$ allows the sum to be normalized between -1 and 1. The search for local maximum of the SVF function gives the instants of important variation of acoustic features. This method [2] has been applied to reduce the number of parameters and mel-cepstrum vectors in automatic

[1] Here a region is defined as an interval centered at this instant, and modifications are carried out on this interval.

speech recognition. We accepted this method which indicates regions of high concentration of acoustic cues because it allows 82% of sound boundaries placed by an expert to be detected. The last 18% are either marks not well placed (at more than 20 ms) or insertions. This second kind of error is not too critical, and even, may be useful ; indeed, the fact of considering more marks, and of slowing down speech in the vicinity of these marks is not likely to reduce the intelligibility. The choice of the slowing down rate is arbitrary, but too strong a value (more than 3) changes the nature of the sound compared to the usual articulation (in particular bursts are artificially transformed into fricatives). We thus accepted a value between 1.8 and 2. Even with this rather high value of slowing down rate, the global average lengthening (for the whole sentence) is only 1.3. It is possible to supplement slowing down by the enhancement of transitory events to improve the perception of some given acoustic cues.

In the first part of this paper, we describe the enhancement strategy which relies on the detection of unvoiced stops and fricatives. In the second part, we present the evaluation of the whole transformation, and, in particular, of the detection algorithm. Lastly, we conclude on the results obtained and future works to be carried out.

2 DETECTING ACOUSTIC EVENTS TO BE ENHANCED

The search for segments to be enhanced mainly focuses on stop bursts and fricatives. According to V. Hazan and A. Simpson [3], the enhancement of this kind of phoneme allows intelligibility of fluent speech to be improved. We decided to focus onto unvoiced stops and fricatives because these sounds can be located with a high degree of robustness and because enhancing these segments improves the perception of the temporal structure of speech.

We explain now how fricatives and stops are detected by using energy criteria. Differentiating a fricative from another sound can be easily achieved on the basis of energy criteria. In first approximation, energy of fricatives is mainly localised in high frequency. We thus chose to calculate the ratio of energy in the band $3600 - 6000$ Hz over that in the band $600 - 1000$ Hz, other frequency bands being considered to be less relevant. Unvoiced fricatives correspond to high values of this ratio.

Stops are characterized by the absence of energy during the closure : this corresponds to a weak average energy at the center of closure segment compared against energy at boundaries. Thus we calculate the average energy [2] at the center of closure, then compare it with the average energy of the whole segment. In the presence of a closure, the average at the center is weaker than the global average. The second cue, which differentiates a burst from a voicing onset, for instance, is the derivative of energy. A burst presents a peak due to the sharp variation of the spectrum at the time of the explosion. A threshold is used to select the significant peaks (50% of the maximum derivative amplitude). The diagram **??** summarises the strategy for detecting bursts and fricatives.

The signal energy enhancement strategy is based on the experiments of V. Hazan and A. Simpson [3]. The principle is to gradually amplify transitions of bursts or fricatives up to a desired level and then to return on the initial sound level at the end of the phoneme.

3 EXPERIMENTS

As mentioned in the introduction the two possible applications are hearing aids and oral comprehension of a second language. At first, we tested oral comprehension (of English sentences by French people) improvement because the adjustment of the strategy to control slowing down and enhancement is substantially easier in this case.

Considering that our transformations are related to perception studies which employ exaggerated acoustic cues [4] we decided to evaluate the improvement at the word level rather than in very specific identification tasks (VCV for instance).

The test corpus consists of 50 sentences selected out of the TIMIT database. 25 sentences were left untouched and the other 25 were modified according to the method presented above.

3.1 Evaluation of transformations

The first evaluation is about the relevancy of the selective slowing down and enhancement. All the 50 sentences of the test corpus were transformed and evaluated. We found only two extra artificial bursts, one of them is masked by the neighboring fricative (and does not change the perception of the word) and the second is perceived as a "click". We observed one pitch marking error at the voicing onset of an unvoiced stop /d/ which changes the scope of the signal amplification and alters the perception. As expected SVF marks appear in regions which contain rapid spectral changes (mainly formant transitions and nasal boundaries). Generally several marks are detected in regions corresponding to rapid formant transitions. As these marks are very close together they give rise to a unique slowing down and do not perturb the slowing down strategy.

3.2 Evaluation of detections

As shows Tab. 1, the detection of the bursts and fricative gives good results (86,8% resp. 88,6%)[3]. These

[2] The frequency band ranging between from 0 and 600 Hz is dismissed from calculation because it could bias the calculation.

[3] The Fig. 1 presents only a simplified description of the algorithm.

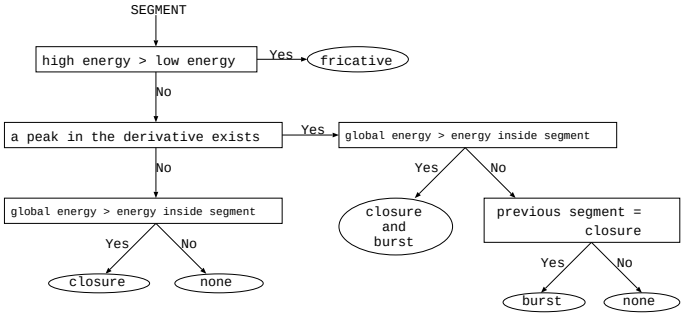

Figure 1: Algorithm of burst and fricative detection

	Bursts	Fricatives
recognized/theorical	177/204(86.8%)	156/176(88.6%)
total number of detections	215	171
correct detections	177(82.3%)	156(91.2%)
omissions	27(13.2%)	20(11.4%)
insertions	38(17.7%)	15(8.8%)

Table 1: Overall results of detections

results were obtained by evaluating, by hand, the detection quality of the events sought by our algorithm. Over the 50 sentences, we counted the effectively perceptible events (bursts and fricatives), the correctly recognised events and analysed the phenomena of omissions and insertions.

First, among the 38 insertions observed during the detection of bursts, 11 can be explained by the quality of the speech. The speaker produces glottal clicks or, for example, the speaker finishes the sentence in guttural manner(on that occasion, one /l/ was detected as a burst). In one sentence, the algorithm detected 3 /n/ and one /m/ not perceptibly marked because the threshold of the 50% of energy derivative is too low for this signal (this is the only case where the threshold is apparently too low). 21 detections (out of the 38) occurred on accentuated phonemes such as /th/ (pronounced as a /t/ or /d/), /n/ or /g/...which can be categorised as plosives, given the pronunciation realised by these speakers. These insertions do not endanger the enhancement from a perceptive point of view and do not deteriorate intelligibility. It should be noted that V. Hazan and A. Simpson [3] developed amplification methods applied to consonants including /n/ which improved intelligibility. The last six insertions correspond to the detection of /f/, this particular case being be discussed below.

Insertions observed during the detection of fricatives, are mainly due to the pronunciation or the addition by the speaker of intense breaths links. Some coarticulation phenomena caused the detection of a fricative segment during bursts (which, besides, would have been forgotten as burst by the algorithm). These last insertions do not matter since the enhancement will be done on the events which contribute to the improvement of intelligibility.

In addition, a certain number of omissions, for the detection of bursts as well as fricatives, comes from errors of the SVF marking algorithm. Thus, three types of errors occurred: either marking split one phoneme into several segments (that leeds to detection failure), or the segmentation of a phoneme was too broad (i.e. the segment includes a part of the adjacent phoneme), or the burst (or the fricative) was in a voiced region, or at least a region, where SVF marks have been put, and thus labeled as a voiced region which prevents the detection of an unvoiced burst or fricative. These errors due to SVF marking account for only 30% of omissions for bursts and 20% for the fricatives. Moreover, these errors cannot be easily avoided since they are directly related to the SVF algorithm.

With regards to the detection of bursts, the other category of omissions originates in the non-detection of the explosion (and thus of the burst) which is usually characterised by a significant peak of the energy derivative. Six cases stem from coarticulation phenomena (due to the fact that left and right phonemes share the same place of articulation) which weaken the explosion and consequently the peak, which goes under the threshold of detection. Ten other cases present too weak a peak in the derivative of energy (even so the closure can be detected). These 16 cases strengthen our choice of accepting a ralative threshold (50%) on the derivative of energy. Even if omissions are possible it is a good compromise to avoid the explosion of the number of insertions if a lower threshold were retained to detect these phonemes. This choice is, therefore, in a perfect adequacy with our strategy of enhancement which consists to only enhance phonemes which do not deteriorate intelligibility after amplification (the counterpart is the risk of omitting some occurrences of these phonemes).

Finally, there are very few fricative omissions (only 8, except errors of SVF marking). On the other hand, it

appears that 6 omissions (out of the 8) are /f/ phones. Consequently, only 3 /f/ out of the 16 (contained in the 50 sentences) were detected as a fricative. We noticed that the acoustic structure of /f/ differs from the other fricatives : the noise of friction is uniform and weak, and lower friction limit can go rather low in frequency. Therefore, the detection of /f/ by our algorithm is very difficult with the current choices made to prevent these false alarms.

In brief, the detection of fricative does not pose any problem (except in the case of /f/). The algorithm is more "strict" for detection of bursts but a good compromise seems to be found between the omissions and insertions which, mainly, do not endanger the intelligibility during the stage of amplification.

3.3 Perceptive evaluation

13 French adults participated in two half-hour experimental sessions. In the first, the 50 sentences were left untouched. In the second session 25 sentences were untouched and the other 25 were modified according to the method presented above. The test corpus was randomized and the subjects were asked to complete one or two missing words in the transcription of the sentence they listened to. We consider four levels of response (none, false, half the phonemes correct, correct).We calculate the average difference (over the 13 listeners) for the 37 words tested between the original and modified versions. A T-test showed that the identification rates improved significantly ($p < 0.02$). A further examination of results showed that improvements are equally distributed between stops and fricatives and that slowing down transitions did not change the perception of transition but has altered the manner of articulation of one burst which has been perceived as a fricative.

4 CONCLUSION

The main strength of our approach is that modifications are carried out only on regions or phonemes which have an important effect on comprehension : transitions - regions with high concentration of acoustic cues - and phonemes like stops and fricatives. The combination of selective slowing down from marks SVF and acoustic enhancement of bursts and fricatives improves the intelligibility of the speech. Each technique gives good results separately and is robust against errors. Indeed, errors made by SVF marking mainly consist of insertions which do not compromise the intelligibility improvement. In the same way, errors made during burst and fricative detection are mainly omissions of weak bursts or fricatives, which are not likely to deteriorate speech intelligibility, unlike insertions (due to false detections) which would introduce artificial bursts or fricatives. Results can be found at *http://www.loria.fr/∼ colotte* [4].

[4]They may be modified following our perceptive tests.

The advantage of our approach is that it is completely automatic ; therefore, it could be easily combined with a synthetic visual speech system in order to supplement the acoustic signal with lip movements in order to exploit the McGurk effect [7].

References

[1] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo. Improved connected digit recognition using spectral variation functions. In *Proc. of Int. Conf. on Spoken Language Processing 1992*, pages 627–630, Banff, Canada, 1992.

[2] G. Flammia, P. Dalsgaard, O. Andersen, and B. Lindberg. Segment based variable frame rate speech analysis and recognition using a spectral variation function. In *Proc. of Int. Conf. on Spoken Language Processing 1992*, pages 983–986, Banff, Canada, 1992.

[3] V. Hazan and A. Simpson. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication*, 24:211–226, 1998.

[4] D. G. Jamieson. Techniques for training difficult non native speech contrats. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, volume 4, pages 100–107, Stockholm, Sweden, 1995.

[5] Y. Laprie and V. Colotte. Automatic pitch marking for speech transformations via td-psola. In *IX European Signal Processing Conference*, Rhodes, Greece, 1998.

[6] S.A. Liu. Landmark detection for distinctive feature-based speech recognition. *J. Acoust. Soc. Am.*, 100(5):3417–3430, November 1996.

[7] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 246:745–746, 1976.

[8] A. Nakamura, N. Seiyama, A. Imai, T. Takagi, and E. Miyasaka. A new approach to compensate degeneration of speech intelligibility for elderly listeners. *IEEE Trans. on Broadcasting*, 42(3):285–293, September 1996.