Perceptual experiments on enhanced and slowed down speech sentences for second language acquisition

Vincent Colotte, Yves Laprie and Anne Bonneau

LORIA/UHP/CNRS
615 rue du Jardin Botanique, BP 101, F-54600 Villers-lès-Nancy, France
{colotte,laprie,bonneau}@loria.fr

Abstract
This paper investigates the perception of speech signals that have been enhanced and slowed down selectively, with the view of improving oral comprehension for second language acquisition.

Our modifications are applied on a small number of acoustic cues, i.e. bursts of unvoiced stops, unvoiced fricative noises and rapid spectral transition regions. Bursts and friction noises were amplified, and spectral transitions were amplified and slowed down. We exploit energy and spectral criteria to localize bursts and friction noises, and spectral variation function to spot rapid transitions.

The perceptual experiment involved students who learn French as a foreign language. The subjects were asked to fill in gaps in incomplete transcriptions of 50 French sentences. The average identification rate increases from 72% up to 81% when the enhancement is applied alone, and up to 86% when the two modifications are applied simultaneously. The strengths of our approach are the robustness of acoustic cue detection and the fully automatic strategy.

1. Introduction
The context of this work is the enhancement of the oral comprehension for second language acquisition. More precisely, our objective is to develop tools to improve speech intelligibility. These tools may operate at the sentence or at the phonetic level.

Nakamura et al. [1] developed a system which slows down the beginning of a sentence and accelerates the end. Their approach takes advantage of the structure of Japanese sentences in broadcast news and is therefore very specific to Japanese. We are convinced that applying transformations at the phonetic level guarantees a relative generality of the strategies, i.e. their independence with respect to the target language.

Several kinds of modifications can be applied: noise reduction, slowing down, local or global enhancement. Our approach originates in our studies on the perception of acoustic cues [2]. We thus are interested in amplifying very specific cues, i.e. burst of unvoiced stops, friction noise of unvoiced fricatives and unvoiced transitions. Furthermore, we want to combine this amplification with a selective slowing down applied in the regions with a high concentration of acoustic cues, i.e. regions where the spectral variations are maximal.

Such tools have been studied in the field of hearing aids to make the distinction between CVs easier for children with hearing deficiencies [3]. In the domain of second language acquisition, Hazan [4] reported improvement in the identification of CV by amplifying bursts, fricatives and nasals.

The originality of our work is to focus on the automatic character of these modifications and their robustness. In fact, we consider that transformations should be piloted automatically and be applied while guaranteeing that they do not introduce any acoustic artifact which could prevent the correct identification of sounds. In the same way, we are convinced that evaluations have to be conducted on automatically modified speech signals in order to measure the overall identification gain.

That explains that we accepted a very small number of acoustic cues (bursts of unvoiced stops, friction noise of fricative and rapid spectral transitions). As these cues are easily detected from speech, and, more importantly, with a high degree of robustness, we do not run the risk of spotting false alarms and introducing acoustic artifacts.

The issue of the strategy for using these tools is crucial because it concerns the acoustic behavior of the transformation of speech signals (i.e. the robustness of the acoustic cue determination and the absence of acoustic artifact) as well as the gain in the identification of speech sounds. The perception experiment presented in this paper is intended to evaluate the relevancy of the strategy of our speech enhancement method.

In the next section we describe the method developed to slow down speech signals selectively and amplify acoustic cues mentioned above. Then we present the perception experiment in the context of learning French as a second language. Finally we present and discuss results obtained.

2. Description of the speech transformation method
The strategy we accepted consists in amplifying bursts, friction noises and unvoiced transitions and slowing down rapid spectral transitions. The implementation aspects cover the detection of these acoustic cues and the slowing down itself.

2.1. Detection of acoustic cues
The localisation of rapid spectral transitions relies on a method which assesses acoustic variations of speech proposed by Flamia [5] and Brugnara [6]. This method, called Spectral Variation Function, uses mel-cepstrum analysis. A coefficient, based on the correlation between spectral vectors in a signal window of 20 ms (with a time shift of 10 ms) and those of neighbouring windows, reflects the spectral variation rate. This method has been originally applied to reduce the number of parameters and mel-cepstrum vectors in automatic speech recognition. We chose this method which indicates regions of high concentration of acoustic cues because it allows 82% of sound bound-
aries placed by an expert to be detected. The last 18% are either marks not well placed (at more than 20 ms) or insertions. Insertions do not introduce any acoustic artifact because they appear in the vicinity of other marks, which only lengthens the scope of slowing down. 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.

Bursts of unvoiced stops and frication noises are detected by using energy criteria. Indeed, differentiating a fricative from another sound can be easily achieved on the basis of energy criteria. In first approximation, energy of fricatives is mainly another sound can be easily achieved on the basis of energy value of slowing down rate, the global average lengthening (for phoneme (or the unvoiced transition).

then to return on the initial sound level at the end of the experiment : eleven Chineses , two Koreans, one Lybian, three Mauritians, one Russian, five Syrians, one Moroccan, two Turks, one Brazilian, one Hungarian. There was about same three Mauritanians, one Russian, five Syrians, one Moroccan, the experiment : one seemed to have some hearing problems, two concerning daily life. Four students were removed from the experiment : one seemed to have some hearing problems, two have too good a French level and one has too bad a French level (the three latter would deteriorate the homogeneous level of this student group). Actually, twenty-eight students participated in the experiment : eleven Chinese , two Koreans, one Lybian, three Mauritians, one Russian, five Syrians, one Moroccan, two Turks, one Brazilian, one Hungarian. There was about same number of female and male subjects. They were about 25 years old in the average (from 21 to 41 years).

The subjects listened to the sounds using headphones MB QUART KN800 in a quiet room (audio room with several separated places). The volume was adjusted to a comfortable level in order to allow all the sentences to be perceived. A weak noise was present but did not disturb listening. The tape with stimuli was recorded from a computer.

The 50 sentences have been selected from the French database DBSONS [9]. These sentences were phonetically balanced and they had been uttered by two speakers (twenty-five by a male and twenty-five by a female). The experimental task consisted in listening to a sentence and filling in one or two gaps in a written incomplete transcription of the sentence. The 50 sentences presented in three different orders for each condition so that one third of the students were given the same order of presentation. All the sentences were played in one session, the subjects were not allowed to stop the tape. A jingle announced each sentence and the sentence was repeated twice. There were three seconds before the beginning of the sentence. There was a nine-second interval between the sentence and its repetition and also after the repetition. There were three example sentences before the 50 sentences of this experiment in order to familiarize listeners. The experiment was about 25 minutes long. For the 28 subjects:

- 9 listened to the 50 original sentences (without any modification) : this is the condition A.
- 9 listened to the 50 sentences which have been enhanced (unvoiced fricatives, stops and SVF transitions) : this is the condition B.
- 10 listened to the 50 sentences which have been enhanced and slowed down (unvoiced and voiced SVF transitions for the slowing down) : this is the condition C.

Modifications - if present - have been applied over the whole sentence and not only over the words to be recognized. Nevertheless, it is possible that some words, corresponding to the gaps to be filled in, have not been enhanced because the acoustic cues have not been spotted properly. But, when the signal has been modified by enhancement and slowing down together, gap words have at least necessarily been slowed down. At last, the words to be identified correspond to 114 sounds (27/p/, 38/t/, 17/s/, 7/f/, 22/k/, 3/f/) distributed over the 50 sentences. Only 106 among these 114 sounds have been actually enhanced : 26/p/, 36/t/, 16/s/, 5/f/, 21/k/, 2/f/. Here is the total number of answers for all the sounds which are considered under each condition :

condition A : 954 (106 × 9) answers
condition B : 954 (106 × 9) answers
condition C : 1060 (106 × 10) answers

4. Results

4.1. Evaluation of transformations

The technique of synthesis gives a high quality signal. The modifications are not perceived by the listeners as a synthetic speech.

Our algorithm detected 92 sounds out of 114 sounds (/p/, /t/, /s/, /l/, /k/, /f/) contained in words to be identified. These omissions are due to some errors of classification of unvoiced and voiced parts (the detection being only succeed on unvoiced parts) and the compromise, we made, between the robustness and the apparition of false alarms. The goal is to enhance only phonemes which do not deteriorate intelligibility after amplification, the counterpart being the risk of omitting some occurrence of these phonemes.
Actually, as the unvoiced transitions are also enhanced, there are 14 other sounds which can be considered as enhanced.

That is why we make a difference between the number of sounds which are theoretically enhanced under the condition B or C and the sounds really enhanced.

4.2. Discussions

The average identification rate (see table 1) was 72% under condition A (no modification), 81% under condition B (unvoiced stops, fricatives and transitions were enhanced) and 86% under condition C (the slowing down of the speech rate is added to the enhancement). Student tests showed that the effects of the modifications (we compared A with B, and A with C) were statistically significant (see table 2). The slowing down of speech rate, when added to the enhancement, only led to a marginally significant improvement ($p = 0.05$) when we took into account all the unvoiced stops and fricatives (actually enhanced or not by our automatic procedure) but was not significant ($p = 0.08$) when we considered only the actually enhanced consonants. We can draw two major conclusions from these results. Firstly, the enhancement alone (condition B), and the enhancement plus the slowing down of the speech rate (condition C) clearly improved the identification of natural stimuli (condition A). Secondly, although the improvements (and their levels of significance) due to condition C were greater than those due to condition B, the addition of the slowing down to the enhancement did not improve in a clearly significantly manner the effect of the enhancement alone. We nevertheless believe that a greater number of data and a better application of our modifications (see the perspectives section) can lead to a clearly significant improvement. Furthermore, the slowing down of the speech rate will probably be of greater interest for vocalic sounds (their study was not in the scope of this paper).

5. Conclusion

The first strength of our approach is that modifications are carried out on regions or sounds which have an important effect on comprehensibility: transitions - regions with high concentration of acoustic cues - and phonemes like stops and fricatives. Furthermore, we do not run the risk of adding ambiguities between unvoiced and voiced consonants because the amplification move unvoiced consonants away from consonants by increasing the distance between these two classes of sounds in terms of energy.

The second advantage is that the modifications are fully automatic and “adjustable”. Therefore, it would be easy to supplement and improve our detection algorithms with techniques from automatic speech recognition.

The signals were modified by means of the PSOLA method which requires pitch extraction and marking. Pitch extraction

Actually, as the unvoiced transitions are also enhanced, there are 14 other sounds which can be considered as enhanced. That is why we make a difference between the number of sounds which are theoretically enhanced under the condition B or C and the sounds really enhanced.

4.2. Discussions

The average identification rate (see table 1) was 72% under condition A (no modification), 81% under condition B (unvoiced stops, fricatives and transitions were enhanced) and 86% under condition C (the slowing down of the speech rate is added to the enhancement). Student tests showed that the effects of the modifications (we compared A with B, and A with C) were statistically significant (see table 2). The slowing down of speech rate, when added to the enhancement, only led to a marginally significant improvement ($p = 0.05$) when we took into account all the unvoiced stops and fricatives (actually enhanced or not by our automatic procedure) but was not significant ($p = 0.08$) when we considered only the actually enhanced consonants. We can draw two major conclusions from these results. Firstly, the enhancement alone (condition B), and the enhancement plus the slowing down of the speech rate (condition C) clearly improved the identification of natural stimuli (condition A). Secondly, although the improvements (and their levels of significance) due to condition C were greater than those due to condition B, the addition of the slowing down to the enhancement did not improve in a clearly significantly manner the effect of the enhancement alone. We nevertheless believe that a greater number of data and a better application of our modifications (see the perspectives section) can lead to a clearly significant improvement. Furthermore, the slowing down of the speech rate will probably be of greater interest for vocalic sounds (their study was not in the scope of this paper).

5. Conclusion

The first strength of our approach is that modifications are carried out on regions or sounds which have an important effect on comprehensibility: transitions - regions with high concentration of acoustic cues - and phonemes like stops and fricatives. Furthermore, we do not run the risk of adding ambiguities between unvoiced and voiced consonants because the amplification move unvoiced consonants away from consonants by increasing the distance between these two classes of sounds in terms of energy.

The second advantage is that the modifications are fully automatic and “adjustable”. Therefore, it would be easy to supplement and improve our detection algorithms with techniques from automatic speech recognition.

The signals were modified by means of the PSOLA method which requires pitch extraction and marking. Pitch extraction
could be a weak point in our transformation, especially if the speech were corrupted by noise. From this point of view, phase vocoder techniques represent an interesting alternative because they allow a vast range of modifications.

Our experiments showed that the improvement is significant for all the sounds. But, due to the small number of occurrences for each consonant, it was not possible to obtain more precise conclusion on the separate contribution of slowing down and enhancement.

Modifications exploited in this work are applied globally, i.e. the whole burst, frication noise or transition is transformed. Besides this work, we are working on the design of acoustic cues for stop identification. We have the project of applying very specific modifications that could selectively enhance acoustic cues involved in the identification of sounds.

Stimuli can be found at [http://www.loria.fr/~colotte](http://www.loria.fr/~colotte).

### 6. Acknowledgements

This work was supported by the Plan Etat-Region Lorrain ("pôle Intelligence logicielle"), project "Téléopérations et assistants intelligents".

We would like to thank Mr. Cambalo and the staff of the Yves Chalon Language Center for providing the listening equipment, and Mr. Trompette for encouraging students of French as a foreign language to participate in this experiment.

### 7. References


