ANALYSIS OF THE DEGRADATION OF FRENCH VOWELS INDUCED BY THE TD-PSOLA ALGORITHM, IN TEXT-TO-SPEECH CONTEXT

Christophe J. Blouin & Paul C. Bagshaw
France Télécom R&D

ABSTRACT

In concatenative speech synthesis systems, synthetic speech is obtained by concatenating acoustic units selected from a database of natural speech. The duration and fundamental frequency (F0) of the selected units are usually different from those requested by a prosodic model, and so some prosodic modification must be applied to the units in order to obtain the desired target. TD-PSOLA is an effective and widely used prosodic modification algorithm, but its use can degrade the perceived quality of the synthetic speech signal. This paper focuses on the evaluation of the degradation of French vowels and determines the influence of several parameters through an analysis of variance. The results show that vowels divide into two groups, based on their first formant frequency (F1). Finally, a modification cost function representative of the degradation is derived from the investigation.

1. INTRODUCTION

The TD-PSOLA (Time-Domain Pitch Synchronous Overlap and Add) algorithm [2] was introduced in a context of diphone synthesis, where the acoustic database contains only one instance of each diphone. Prosodic modifications (duration and fundamental frequency changes) of the units are necessary to match the prosodic target given by a prosodic model. Since the TD-PSOLA algorithm distorts the speech signal, larger databases containing several instances of each unit may be used to minimise the extent of the modifications, and thus minimise the induced distortion. However, the unit selection process does not always succeed in creating natural-quality synthetic speech because there are never sufficient well-suited units in the database, and prosodic modifications remain necessary.

One way of selecting the best-suited segment from the database is to define a cost function, generally a linear combination of a target cost and a concatenation cost [4]. The target cost represents the distance between the target unit requested by the prosodic model and candidate units, and the concatenation cost represents the distance between two adjacent selected units. The prosody modification algorithm is applied at the final stage of the synthesis, once all the units to concatenate have been chosen. However, the use of a prosodic modification algorithm is related to the target cost, which should take into account the perceptual distortions induced by the prosodic modification algorithm.

Only a few studies have yet been conducted to evaluate the effect of the prosody modification by TD-PSOLA. Moulines & Charpentier [8] report the traditional limitation for duration modification as a factor of 2. Hirokawa & Hakoda [3] study the F0 modification in a global way on three sentences, and determine the F0 modification limits as 80% to 125% for an acceptability rate of 75% (13 subjects). Kawai et al. [6] use an approach comprising both F0 and duration modifications, which leads to global F0 modification limits of 66% to 132% and global duration modification limits of 57% to 115% for an opinion score of at least 4 out of 5 (5 subjects). In a psychoacoustical test (3 subjects), Kortekaas & Kohlraush [7] modify synthetic two-formant signals and show that F1 is important for the perception of small F0 modifications. This observation cannot be used directly in the current speech synthesis context because of the synthetic nature of their test stimuli.

The aim of this study is to evaluate precisely the degradation induced by the TD-PSOLA algorithm and to define a modification cost function and modification limits for each French vowel. The paper is organised as follows. Section 2 describes the evaluation procedure and the stimuli presented to the subjects. In section 3, the results of the evaluation are analysed through an analysis of variance of several parameters, and the influence of the parameters is outlined. In section 4, a modification cost function is derived from the data collected during the test. To conclude, the results of this study and future work are discussed.

2. LISTENING TESTS

2.1. Test Procedure

To evaluate the degradation introduced in the speech signal by the TD-PSOLA algorithm, listening tests were conducted by 16 subjects, following the ITU-T P.800 recommendations [5]. A DCR rating procedure was conducted because of its sensibility to subtle degradation. The subjects (8 men and 8 women) were all French natives, under 35, naive in the sense of perceptual listening tests and with no known hearing impediments. They were paid to participate in the tests. The 625 stimuli are composed of disyllabic non-sense words (judging the degradation in monosyllabic stimuli was found be too difficult).

The listeners were presented with pairs of stimuli, composed of a natural reference, a 500ms silence, and a prosody-modified version of the natural reference. Between pairs, the listeners had 5secs to judge the degradation due to the modification. The tests were divided into twelve 10-minute sessions. Stimuli such as the vowels used by Bailly [1] (a natural target and a natural source) would have been ideal, but are not used here because such vowels are isolated and have artificially long durations.

The subjects were asked to judge only the degradation introduced in the second stimulus of each pair, disregarding the prosodic modifications, and to rate it on a scale from 1 (very disturbing degradation) to 5 (imperceptible degradation or even...
improvement). The subjects conducted two short training sessions, each of 15 stimuli, representative of the entire degradation range, in order to familiarise them with the evaluation procedure before the tests.

2.2. Stimuli Description

Contemporary French language now tends to differentiate only 13 phonemic vowels. These 13 French vowels are evaluated, namely 10 oral vowels (/a/, /e/, /i/, /u/, /o/, /ø/, /y/, /ɛ/, /ɛ/, /ɛ/ and /ɛ/) and 3 nasal vowels (/ɔ/, /ɑ/ and /ʊ/). The stimuli non-sense words are of the form CVVCV where the two consonants are unvoiced plosives. These contexts ensure that the vowels’ phonemic targets are reached, by minimising co-articulation.

The natural references were uttered by a professional male speaker, whose voice is also used in the synthesis system’s database. The speaker has a natural low voice with an average F0 of 90Hz. The reference utterances are pronounced at a steady rate and with a neutral tone and regular pitch.

The prosody-modified versions are generated from the natural utterances with the following possible modifications attributed to one of the vowels. Two F0 shapes are possible, one “uniform” and the other “dynamic”. For the “uniform” F0 shape, F0 is kept constant at 50, 75, 100, 150 and 200% of the original mean F0. For the “dynamic” F0 shape, F0 is linearly varied from respectively 50, 75, 100, 150 and 200% to respectively 200, 150, 100, 75 and 50% of the original mean F0. The duration modifications are 50, 75, 100, 150 and 200% of the original duration. The large range of modifications in the stimuli ensures that some degradation is introduced and that the limit of acceptable modifications can thus be determined.

Stimuli are generated with each of the 13 vowels in the first syllable, uttered with a normal tone and modified by 5 “uniform” F0 modifications and 5 duration modifications (325 stimuli). Stimuli for 4 vowels (/ɛ/, /ɛ/, /ɛ/ and /ɛ/) are also uttered two additional times, once in the second syllable (normal tone) and once with a high-tone (in the first syllable). They are, once again, modified by 5 “uniform” F0 modifications and 5 duration modifications (200 stimuli). An analysis of these high-tone vowels showed an average difference of 20% of the original F0, which corresponds to a variation of roughly 25% of the speaker’s F0 range. For these same 4 vowels, in first syllable position and uttered with a normal tone, the natural utterances are also modified by the 5 “dynamic” F0 modifications and 5 duration modifications (100 stimuli).

3. ANALYSIS OF THE RESULTS

Two training sessions were necessary for the subjects to assess coherently the degradation, and even with these two sessions, the judgements of roughly the first 50 stimuli of the test had a larger standard deviation than judgements of the remainder of the test.

An analysis of variance (ANOVA) was performed on the subjects’ judgements of the degradation in the vowels. The statistics software used was not able to analyse simultaneously all the data collected, thus, the analysis was conducted on several groups of stimuli and their related judgements. These groups were formed as a function of the parameters studied. The parameters used in this analysis are: vowel type, VP, nasal or oral character, ON, vowel position in the word, VP (1st or 2nd syllable), original pitch, OP (normal or high), the F0 shape, FS (uniform or dynamic), degradation modification, DM and F0 modification, FM. The results were considered as significant (p<0.05), highly significant (p<0.01) or extremely significant (p<0.001).

3.1. Effect of the Oral Vowel Type

Considering the 10 oral vowels, the parameters VT, FM and DM are studied (OP = normal, ON = oral, VP = 1st, FS = uniform).

An extremely significant effect of DM (F=53.316; df=2.052; p<0.001) was found, characterised by a strong increase of the degradation when increasing the duration, cf. figure 1: Degradation Mean Opinion Score (DMOS) against duration modification, (all vowels and all F0 modifications averaged). The other significant effects are those of VT (F=9.580; df=0.956; p<0.001) and of FM (F=8.038; df=10.506; p<0.001), cf. figure 2.

All the interactions concerning the oral vowels are significant but are weak, compared to the direct effects of the parameters. The interactions, in order of significance are: the one between VT and DM (F=1.881; df=0.512; p<0.01), the one between VT and OP (F=1.235; p<0.05), the one between VT and FS (F=9.254; df=0.673; p<0.05), the one between VT and FS and DM (F=3.85; df=0.524; p<0.01), the one between VT and OP and FS (F=2.859; df=0.698; p<0.001), the one between VT and DM and OP (F=1.181; df=0.524; p<0.01), the one between FM and VT (F=1.509; df=0.559; p<0.05), and the one between the three parameters (F=1.503; df=0.484; p<0.001).

These results show that the effect of DM has the main influence and that the parameters DM and FM are weakly related. The set of the vowels divides into two significantly different groups: the vowels /ɛ/, /ɛ/ and /ɛ/ obtain a DMOS on average 0.30 above the others. Note that these three vowels are the only ones for which the first formant frequency is below 250Hz.

3.2. Effect of the Nasal Vowel Type

The nasal vowels (/ø/, /ø/ and /ø/) are compared with their respective oral equivalents (/ɛ/, /ɛ/ and /ɛ/). Four parameters are evaluated in this case: ON, VT, DM and FM (OP = normal, VP = 1st, FS = uniform). For these six vowels, the main significant effect is that of DM (F=44.707; df=1.737; p<0.001), followed by effect of FM, (F=9.029; df=7.798; p<0.001) and then that of ON (F=6.329; df=1.235; p<0.05). There are several interactions: the most important amongst them is the one between ON and FM (F=12.727; df=0.606; p<0.001). The others are: between DM and FM (F=4.604; df=0.573; p<0.001), between VT and DM...
Figure 3

The effect of the vowel type is not significant: this means that vowel degradation is the same in each of these groups. As the interaction between ON and VT is not significant and the effect of ON is weak, oral vowels and their nasal equivalents tend to have roughly the same degradation. However, it can be deduced from figure 3 that the nasal vowels degrade more than their oral equivalents can when lowering their F0 (all DM values averaged). This dissymmetry should be taken in account by the degradation cost function, to avoid lowering F0 in nasal vowels.

3.3. Effect of the Position in the Word

The 4 vowels /a/, /i/, /u/ and /V/ in the first syllable are compared with their counterparts in the second syllable. The parameters VP, VT, FM and DM are studied here (OP = normal, ON = oral, FS = uniform).

The main significant effect is associated with DM (F=79.138; df=1.302; p<0.001). The other significant effects are those of VT (F=29.697; df=2.220; p<0.001), of FM (F=10.829; df=1.302; p<0.001) and of VP (F=7.676; df=1.795; p<0.05). The second order interactions are: between VP and VT (F=23.597; df=1.153; p<0.001), between DM and VT (F=8.599; df=0.497; p<0.001), and between FM and VT (F=4.805; df=0.608; p<0.001). The significant interactions in the third order are: between VP, DM and VT (F=6.642; df=0.483; p<0.001), and between VP, VT and FM (F=3.305; df=0.643; p<0.001).

Figure 4

The parameter VT seems to have a great influence, but the effect of VT and all the interactions are in fact accounted for by the behaviour of the word-final /a/: with almost no degradation for duration increase. The other word-final vowels tend to degrade more than their beginning-of-word equivalents, cf. figure 4.

3.4. Original Pitch

The 4 vowels /a/, /i/, /u/ and /V/ uttered with a high tone are compared with their normal-tone equivalents. The parameters OP, VT, FM and DM are studied (ON = oral, VP = 1”, FS = uniform).

The analysis shows that the main significant effect is that of DM (F=56.062; df=2.042; p<0.001). The effects of OP and of VT have roughly the same importance, respectively (F=14.170; df=3.964; p<0.01), and (F=12.441; df=1.782; p<0.001). The less important effect is that of FM (F=8.771; df=10.117; p<0.001). Interactions between the four parameters are almost all significant, but have no strong influence, and only the most important of them are outlined here. The most important interactions are: between OP and FM (F=4.762; df=0.598; p<0.01), between VT and DM (F=4.562; df=0.557; p<0.001), and between VT and OP (F=3.071; df=1.039; p<0.05).

Figure 5

The results here show that high-tone vowels are on averaged 0.25 DMOS more degraded than their normal-tone equivalents, cf. figure 5 (all DM and FM values averaged). This is in concordance with results in the literature, and this means that not only the average F0 of a speaker must be taken into account during selection process, but also the mean F0 variations within the unit database.

3.5. Type of F0 Shape

Up to this point, the analysis has been concerned with a “uniform” F0 shape. However, the F0 shape required by speech synthesisers is rarely constant, and a more realistic modification needs to be examined. Thus, the use of a “dynamic” F0 shape (linearly varying F0) which is closer to a system’s real requirements. The parameters studied here are: FS, VT (fa, fi, fa and /s/), DM and FM (OP = normal, ON = oral, VP = 1”).

The effect induced by DM is once again the main one (F=50.477; df=1.917; p<0.001), followed by that of FS (F=23.104; df=1.067; p<0.001), of VT (F=19.096; df=1.660; p<0.001), and of FM (F=14.608; df=1.039; p<0.001). The interactions are: between DM and VT (F=8.330; df=0.487; p<0.001), and between DM and FM (F=2.419; df=0.613; p<0.01). Note that there is no significant interaction of second order between the parameter FS and any of the other parameters. The third order interactions are: between FS, VT, and DM.
(F=2.586; df=0.406; p<0.01), and between VT, DM and FM (F=1.445; df=0.477; p<0.05).

It can be deduced that in all cases the “dynamic” modification is preferred to the “uniform” modification. These differences are not always significant as is the case for the average of the 100% F0 modifications, which is in fact identical for both “uniform” and “dynamic” modification. This could be related to the F0 modification of the vowel, which is applied to the stable part of the vowel in the case of “uniform” modification, and only on the unstable part of the vowel for “dynamic” modification. This phenomenon should be taken into account when selecting units in the database, by appropriately weighting the F0 of the stable and the unstable parts of vowels.

4. DEGRADATION COST FUNCTION

A degradation cost function can be derived from the shape of the non-linear vowel degradation curve. This function is composed of three sub-functions, one for each of the three vowel groups identified by the analysis presented in section 3: the group of vowels /i/, /u/ and /y/, the group of other oral vowels, and the group of nasal vowels. These sub-functions have 2 input variables (duration modification and F0 modification) and are obtained by averaging the results for each vowel of a group weighted by the standard deviation of each DMOS.

The modification limit is defined as a function of the quality expected and of the amount of data available, corresponding to a mean judgement between 3 and 4.5. For their use in text-to-speech synthesis, the degradation functions have to be interpolated to predict the degradation for any required prosody and vowel. The “dynamic” F0 shape is not taken into account in the degradation functions, so the predicted degradation corresponds to the worst-case degradation. Thus, no extra degradation can be induced by the modification, whatever the new F0 shape is. The case of high-tone vowels is also not taken into account, because it was found to be neither very important nor significant. The fact that all the vowels tested had reached their phonemic targets will over-estimate the predicted degradation; this ensures that no extra degradation can happen in real synthesis, whatever the vowel’s context is.

5. CONCLUSION

By conducting perceptual tests on the degradation induced by TD-PSOLA, this study investigates the influence of several parameters. The major results of the experiments are:

- The most important first order effect is duration modification, with a strong degradation when increasing the duration.

- Two significant groups of vowel behaviour are found, related to their first formant frequency: low F1 vowels as /i/, /u/ and /y/ obtain a DMOS on average 0.35 above the others.

- F0 modification has a less important influence, and induces the same degradation when increased and lowered. An exception was found with the three nasal vowels, for which lowering F0 induces more degradation than increasing it.

- The F0 shape induced different degradation. This is related to the total F0 modification and to the part of the vowel where the modification is performed.

The results also verify those already reported in the literature. The original pitch is known to have an influence on the voice quality when using TD-PSOLA [8]. The higher the original pitch, the poorer the quality, so female voices are usually more difficult to process than male voices. There is a weak interaction between duration modification and F0 modification, or the influence of the original pitch [8]. The importance of F1 is underlined with the two groups of vowel behaviour, which is in concordance with the results of [7] even if their signals had F1 superior or equal to 500 Hz. There is a resemblance between the degradation cost sub-function defined here and that found by [6].

To conclude, the parameters’ influence and the degradation cost function defined in this study should enhance the selection of units when using the TD-PSOLA algorithm to perform the prosodic modifications. The degradation was studied on French vowels, but the relation with F1 could apply to other vowels. Future work will include listening tests to study degradation induced in consonants.

6. REFERENCES


