ISCA Archive
http://www.isca-speech.org/archive
8th ISCA Speech Synthesis Workshop • August 31 – September 2, 2013 • Barcelona, Spain

Evaluation of contextual descriptors for HMM-based speech synthesis in French

Sébastien Le Maguer[1], *Nelly Barbot*[1], *Olivier Boeffard*[1]

[1]IRISA - University of Rennes 1, Lannion, France
{Sebastien.Le_Maguer, Nelly.Barbot, Olivier.Boeffard}@irisa.fr

Abstract

In HTS, a HMM-based speech synthesis system, about fifty contextual factors are introduced to label a segment to synthesize English utterances. Published studies indicate that most of them are used for clustering the prosodic component of speech. Nevertheless, the influence of all these factors on modeling is still unclear for French.

The work presented in this paper deals with the analysis of contextual factors on acoustic parameters modeling in the context of a French synthesis purpose. Two objective and one subjective methodologies of evaluation are carried out to conduct this study. The first one relies on a GMM-approach to achieve a global evaluation of the synthetic acoustic space. The second one is based on a pairwise distance determined according to the acoustic parameter evaluated. Finally, a subjective evaluation is conducted to complete this study.

Experimental results show that using phonetic context improves the overall spectrum and duration modeling and using syllable informations improves the F0 modeling. However other contextual factors do not significantly improve the quality of the HTS models.

Index Terms: HTS, Evaluation, Contextual factors, French synthesis

1. Introduction

Based on Hidden Markov Models, HTS [1] provides a framework to synthesize speech using parametric statistical models offering a good flexibility. Acoustical parametrization is generally done with a MLSA filter [2] associated with the STRAIGHT model [3]. To produce an acoustic signal for a specific utterance, the temporal evolution of the acoustic parameters is generated from a sentence-level HMM whose observations encompass segmental and prosodic (f0 and duration) informations. This sentence-level HMM is built by concatenating HMM related to the phonemes which compose the utterance.

In HTS, a phone is qualified by a set of contextual factors. For example, the set of describing factors for English, introduced in [4], contains about fifty contextual descriptors associated to the phonetic, phonologic, prosodic or linguistic properties of a segment. During the HTS clustering stage, these factors are used to guide the construction of a decision tree. Consequently, contextual factors have an important role in model training which implies that choosing a proper set could influence the quality of the contextual HMM.

Even though HMM-based synthesis systems are evaluated during the Blizzard challenge [5], only few studies are focused on the influence of contextual factors on the acoustical parameter modeling and, then, on the synthesis achieved by HTS. In [6], the acceleration parameters are studied by computing distances between generated parameters containing acceleration coefficients and those without acceleration coefficients. This study shows that discarding acceleration coefficients implies a saw-tooth trajectory generation. In [7], the duration prediction error is evaluated using RMSE and the correlation between the generated duration and the original duration associated to the same utterance. The results in [7] indicate differences in modeling the duration of consonants and vowels. Specifically to the contextual feature issue, the contribution of high level linguistic features along with the influence of hand labeled versus automatic labeled features are assessed in [8]. This study shows that using automatic annotation in the training labels could affect HTS modeling and that pitch accents, boundary tones and POS (Part-Of-Speech) tags contribute more than other phrase level contextual features to the modeling. By extending this result, we can assume that some contextual features are less effective than others. This assumption is confirmed by the prosodic contextual factor evaluation conducted by [9] to identify a minimal descriptive feature set. Finally, using this assumption, it is possible to achieve "on-the-fly" synthesis like the one proposed in [10].

The aim of this paper is to propose a protocol for an objective evaluation of the HTS synthesis and to apply this protocol to analyze the speech generated by HTS for French. The first method we propose is based on an acoustic space modeling. By analogy to voice conversion, we assume that the acoustic space is well represented by a Gaussian Mixture Model (GMM). By comparing the likelihood of each GMM, which models a generated acoustic space, given a reference speech dataset, it is possible to compare the similarity of the different acoustic spaces. In this way, we can study the quality of the acoustic spaces generated according to different sets of contextual features. In addition, comparing acoustic spaces using a GMM likelihood does not require an alignment between the HTS synthetic speech and the natural reference. Then this approach enables an evaluation of acoustic parameters independently of the duration. However, we need enough data to train the GMM which prevents precise analysis by using this method. Consequently, a second objective methodology, based on usual distances, is used in this protocol for local analysis. During experiments these distances allow to assess the modeling quality according to phonetic categories. In addition, a global subjective MOS test is proposed to compare synthesized speech obtained with different contextual factor combinations and natural speech.

This paper is organized as follows. Section 2 presents the objective evaluation protocol. Section 3 exposes data and the results of the experiments. Section 4 describes the subjective evaluation protocol and its results.

2. Objective evaluation

2.1. General framework

The purpose of the proposed protocol is to study the influence of various descriptors on the acoustic space generated from a single-speaker HTS system, and to assess its proximity to the acoustic space associated with natural speech of the same speaker.

The set of descriptors used to qualify a phonetic segment is the one given by [4] with some adaptations. First, information concerning lexical accent at the syllable level and the TOBI labels at the sentence level are overlooked. Secondly, we used specific French tools to retrieve the POS tags. The descriptors are introduced in table 1. In order to achieve our study, several subsets of contextual factors have been defined. They are presented in table 2.

The acoustic space of the speaker, estimated from STRAIGHT analysis-by-synthesis signals, will serve as a reference. In this specific case, denoted a/s, the HTS system is not used (it is the best case for the objective evaluation experiments). In the following paragraphs, the notations $L_{a/s}$, $V_{a/s}$ and $T_{a/s}$ will refer to three sets of acoustic vectors (corresponding respectively to the Learning, Validation and Test corpora) stemming from analysis-by-synthesis signals, corresponding to disjoint sets of utterances.

For each subset of contextual factors $k \in \{$p1,...,p5-s_pos$\}$, the learning phase of the HTS is done on the $L_{a/s}$ corpus using the k set only. A corpus L_k of acoustic vectors (respectively V_k and T_k) is generated by HTS, corresponding to the same utterances as $L_{a/s}$ (respectively $V_{a/s}$ and $T_{a/s}$).

In order to compare the acoustic spaces generated by HTS with the one based on analysis-by-synthesis signals, two objective evaluation methods are being considered. One is based on a GMM modeling of the acoustic spaces, and the other relies on a distance between the vectors generated by HTS and the vectors stemming from analysis-by-synthesis processing. In order to assess, in an independent way, the quality of each HTS parameter, the duration of the segments observed in $T_{a/s}$ is forced upon the generation process of the elements of T_k (for each $k \neq a/s$ set) in case of the evaluation of MGC (Mel Generalized Cepstral) coefficients and F0 values synthesized by HTS.

2.2. Evaluation based on GMM

This methodology mainly relies on the following assumption: if a configuration of HTS improves the quality of the synthesized speech signal, the likelihood of the reference data with respect to the acoustic space generated by HTS should increase. Since the likelihood depends on both the model and the data, we have chosen to keep the same test corpus $T_{a/s}$ as a referential throughout this evaluation process.

For every $k \in \{$a/s,...,p5-s_pos$\}$, the GMM $\mathcal{M}_k$ is learnt over L_k using an EM algorithm. According to the evaluated HTS parameters, each vector of L_k could correspond to the spectral part of frames, the fundamental frequency of frames or the duration of phones. In case of evaluation of the spectral part, L_k vectors are 39-order MGC coefficient vectors. The 0^{th} MGC coefficient corresponds to the gain and is ignored in order to facilitate the comparison with the evaluation based on mel-cepstral distortion (eq.1).

For each of these data types, the GMM-based evaluation methodology is similar. However, in case of the spectral evaluation, a principal component analysis (PCA) is operated on the whole set of learning corpora in order to reduce the dimension of their elements and ensure the numerical stability during the learning stage of $\mathcal{M}_k$. The target threshold of the PCA is at least 95% of the explained variance in the data. The PCA linear transformation $\mathcal{T}$ is also applied to the vectors of V_k, T_k and $T_{a/s}$ so as to homogenize the data. After the application of the PCA, with no risk of confusion, notations L_k, V_k and T_k are conserved.

The number of components n of the GMM $\mathcal{M}_k(n)$ is determined using the validation corpus V_k: for $i \in [1..9]$, the $\mathcal{M}_k(n)$ model is learnt over L_k for $n = 2^i$ and the log-likelihoods $LL(L_k|\mathcal{M}_k(n))$ and $LL(V_k|\mathcal{M}_k(n))$ are then computed. The covariance matrices of the GMM components are diagonal. Finally, an over-learning situation is detected when $LL(V_k|\mathcal{M}_k(n)) << LL(L_k|\mathcal{M}_k(n))$. The optimal value of n, known as $n^\star$, is then chosen as the minimal number 2^i so that $LL(L_k|\mathcal{M}_k(n)) - LL(V_k|\mathcal{M}_k(n)) \geq \epsilon$, for every $k \in \{a/s, \ldots, p5 - s\_pos\}$, where ϵ was a priori defined to $\epsilon = 0.2$.

The log-likelihoods of the data from the test corpora $LL(T_{a/s}|\mathcal{M}_k(n^\star))$ and $LL(T_k|\mathcal{M}_k(n^\star))$ are then computed, along with the associated 95% confidence intervals using a Bootstrap methodology. This allows for the evaluation of the adequacy of the HTS-generated acoustic spaces with the reference data coming from the analysis-by-synthesis STRAIGHT process.

2.3. Evaluation based on distances

The aim of this methodology is to dispose of a measure that enables a local analysis of the closeness between the coefficients generated by HTS and those stemming from the STRAIGHT analysis.

In case of the evaluation of MGC vectors and F0 values, the segments provided by HTS have the same duration as the STRAIGHT segments and the frames of T_k and $T_{a/s}$ can be matched for each $k \in \{$p1,...,p5-s_pos$\}$.

The measure considered here between two 39-order MGC vectors c_k and $c_{a/s}$, respectively elements of T_k and $T_{a/s}$, is the mel-cepstral distortion expressed as

$$D\left(c_k, c_{a/s}\right) = \frac{10\sqrt{2}}{\ln(10)} \sqrt{\sum_{i=1}^{39} \left(c_k(i) - c_{a/s}(i)\right)^2}. \quad (1)$$

This distortion is computed for all the $(c_k, c_{a/s})$ pairs of $T_k \times T_{a/s}$ and a confidence interval of the associated mean value is also computed for each $k \neq a/s$.

The distance between the F0 and duration values generated by HTS and their matched elements in $T_{a/s}$ is derived using a Root Mean Square error (RMS). More precisely, for each k subset, the RMS error between the F0 frames of T_k and $T_{a/s}$ is in cent. We have used 87 Hz as the reference frequency which represents the mean F0 value of the speaker. For the phone duration, the RMS error is computed taking into account all the phone instances of $T_{a/s}$.

At last, the voicing error rate has been introduced to complete the analysis of the fundamental frequency. This measure is used to analyze specifically the voicing boundary F0 modeling. Considering the F0 values c_k and $c_{a/s}$, respectively elements of T_k and $T_{a/s}$, the voicing error is defined by

$$D\left(c_k, c_{a/s}\right) = \begin{cases} 0, & \text{if } c_k = c_{a/s} = 0 \\ 0, & \text{if } c_k \neq 0 \text{ and } c_{a/s} \neq 0 \\ 1, & \text{otherwise.} \end{cases} \quad (2)$$

	Id.	Horizon	Meaning
Pho.	A		Identity of the current segment
	B		Identity of the previous/next segment
	C		Identity of the previous-previous/next-next segment
Syllable	D	P/C/N	Number of phones + position of the current phone in the syllable
	E	C	Position of the syllable in the word
	F	C	Position of the syllable in the sentence
	G	P/C/N	Accented flag
	H	C	Number of syllables from the (previous accented)/current syllable to the current/(next accented) syllable
	I	C	Number of accented syllable before/after the current syllable in the sentence
	J	C	Vowel of the syllable
Word	K	P/C/N	Number of syllables in the word
	L	C	Position of the word in the sentence
	M	P/C/N	Word POS tag
	N	C	Number of words from the (previous content)/current word to the current/(next content) word
	O	C	Number of content words before/after the current word in the sentence
Sent.	P	P/C/N	Number of syllables in the sentence
	Q	P/C/N	Number of words in the sentence
	R	C	Position of the sentence in the utterance

Table 1: Contextual factors used for French speech synthesis. The second column indicates which items are described (P=Previous, C=Current and N=Next)

		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
Seg.	p1	X																	
	p3	X	X																
	p5	X	X	X															
Syl.	p5-sy_pos	X	X	X	X	X	X												
	p5-sy_accent	X	X	X				X	X	X	X								
	p5-sy_full	X	X	X	X	X	X	X	X	X	X								
Word	p5-w_pos	X	X	X	X	X	X	X	X	X	X	X	X						
	p5-w_content	X	X	X	X	X	X	X	X	X	X			X	X	X			
	p5-w_full	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
Sentence	p5-s_pos	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

Table 2: Evaluated contextual factor sets. An "X" indicates that the factor belongs to the set.

The voicing error rate corresponds to the average voicing error value multiplied by 100.

3. Experiments and results

3.1. Data

The data used for the evaluation are extracted, using a full automatic process presented in [11], from an audiobook in French. The speaker was a male speaker whose reading was moderately expressive. The signal was sampled at 16kHz. The HTS version is the speaker-dependent architecture presented at the Blizzard challenge in 2005[1]. Utterances from $L_{a/s}$ are used to train the HMM models in HTS for each contextual descriptor set k under consideration.

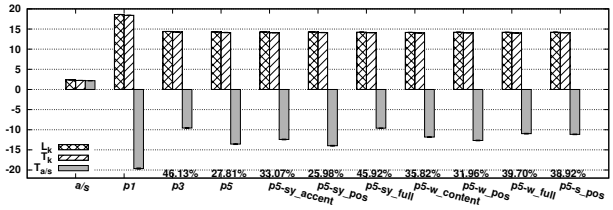
As previously mentioned, three sets of utterances are built: a training corpus containing about 300 utterances for a duration of one hour, a test corpus and a validation corpus which both contain 152 utterances for a duration of 10 minutes. Furthermore, for the two objective evaluations, all frames associated with non speech sound labels (pauses, noises, etc) are simply discarded. Therefore, the training corpus contains about 520,000 frames; the test and validation corpora contain about 85,000 frames each.

3.2. GMM-based evaluation results

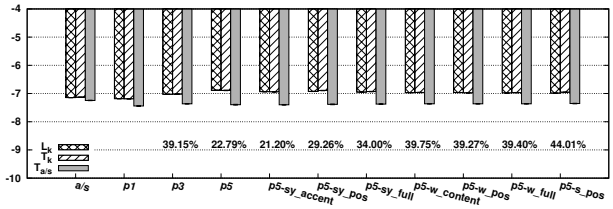
At the end of the GMM learning stage described in section 2.2, the resulting GMM are composed of 512 Gaussians for the spectral part evaluation, 128 for the F0 and 2 for the duration. Furthermore, for the spectral part evaluation, the application of the PCA reduces the data dimension from 39 to 12. Results of the GMM evaluation method are illustrated in figure 1.

For all evaluated acoustic parameters and considering $T_{a/s}$ as a reference, the highest likelihood of its elements is obviously provided by $\mathcal{M}_{a/s}$ and the lowest one by $\mathcal{M}_{p1}$: using only the phonetic label of the current phone segment is not enough to generate an appropriate acoustic space according to the coefficients extracted from the natural speech signal. Globally, by taking into account the closest phonetic context (one left and right phonetic context), the likelihood of the data stemming from analysis-by-synthesis and relative to the GMM generated from HTS acoustic vectors increases significantly.

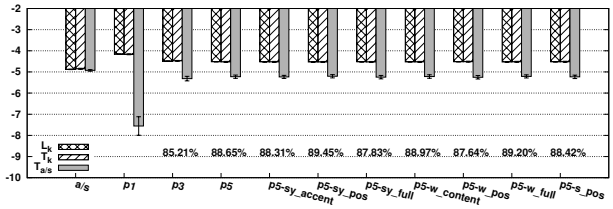
Differences between acoustic parameters appear when more contextual factors are used. As for the segmental part, the best contextual factor set is p3 and, according to the presented results, taking into account more features sometimes could lead to produce more irrelevant data. As for the duration, we can observe that there is a constant improvement until the syllable level. However, concerning the duration, confidence intervals



(a) MGC avoiding the gain dimension



(b) F0



(c) Duration

Figure 1: Log-likelihoods of L_k, T_k and $T_{a/s}$ for GMM $\mathcal{M}_k$, where descriptor combination k is given in the x-axis. Rates below the bars indicate the improvement rates $(LL(T_{a/s}|\mathcal{M}_k) - LL(T_{a/s}|\mathcal{M}_{p1}))/(LL(T_{a/s}|\mathcal{M}_{a/s}) - LL(T_{a/s}|\mathcal{M}_{p1}))$ associated to each k from $p3$ to $p5 - s\_pos$ compared to $p1$.

overlap which means that from p3 to p5-s_pos, all contextual factor sets are equivalent. Finally, results achieved by the evaluation for F0 indicate that all contextual factor sets are equivalent. So, HTS globally provides suitable F0 space with respect to the analysis-by-synthesis data.

3.3. Pairwise evaluation results

3.3.1. Spectral evaluation results

We present results of the second objective methodology based on mel-cepstral distortion between the original spectral coefficients and the ones generated by HTS, are illustrated in figure 2.

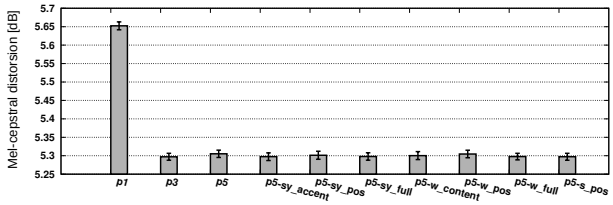


Figure 2: Average mel-cepstral distortion between $T_{a/s}$ and T_k vectors for several combinations of contextual descriptors presented on the x-axis

These results show that speech generated using only the current phonetic label of a segment (p1 set) is farthest from the natural speech signal. This is consistent with the GMM evaluation results. In addition, the lowest distortion is achieved using the set p3 with no significant difference with more complete contextual factor sets. Furthermore, according to the results provided by the GMM based evaluation, this can mean that using some sets, like p5 for example, leads to consider some analysis-by-synthesis values unlikely even if the generated values are not so far from them.

In order to post-analyze potential sources of errors, sets of vectors are defined according to the phonetic characteristics (consonant/vowel, voiced/unvoiced/oral/nasal, etc) and the associated mel-cepstral distortions are given in figure 3.

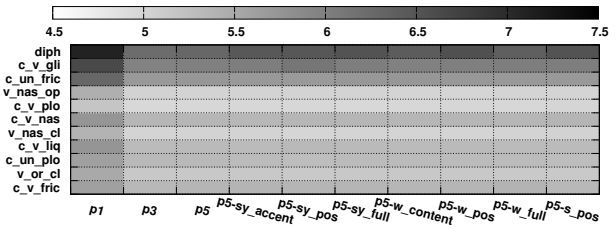


Figure 3: Mel-cepstral distortion presented by phonetic characteristics for each contextual factor set. Distortion, expressed in dB, is quantified on a grayscale. Each set is labeled by x_y_z where x$\in$ {c(onsonant), v(owel)}, y$\in$ {v(oiced), un(voiced), or(al), nas(al)} et z$\in$ {gli(de), fric(ative), (im)plo(sive), op(en), cl(osed), liq(uid), nas(al)}. Diph. set is represented only by the phone /yi/.

The distortion values between analysis-by-synthesis coefficients and generated coefficients based on p1 descriptor set are the highest ones. Confidence intervals, which are not present in this figure, confirm that those differences are significant relatively to other contextual descriptor combinations. By comparing mel-cepstral distortions between the different descriptor sets, we distinguish three main sets: vowels with voiced plosives, diphthongs and unvoiced fricatives, and the other consonants. None of contextual factors introduced in the French set seem to fill the gaps between those main sets. Considering the diphthong, the distortion can be explained by number of frames (about 2.000 frames) used in the training stage but this explanation is not suitable for other phonetic sets (from 7 to 90 times greater). We conclude that none of the contextual descriptors used can really capture the specific acoustic properties of, for example, glides as much as open nasal vowels.

3.3.2. F0 evaluation results

The results obtained by applying the pairwise evaluation on the F0 are presented in figures 4 and 5. Using high-level contextual factors does not improve the error rate. Indeed, the best voicing error rate is achieved by using the direct phonetic context (labels of the previous, current and next segments). However, by comparing the RMS values, we notice a constant improvement until the set p5-sy_full. Taking into account higher level contextual factors implies a statistically significant improvement of the RMS. So, according to those results, the best contextual factor set for the F0 modeling is p5-sy_full.

By comparing these results with the GMM-based evaluation ones, we can notice a clear difference. If we analyze globally the generated values, most of contextual factor sets lead to produce equivalent F0 spaces. However differences between the generated F0 values occur more locally. So, even if the F0 values produced by most of the contextual factor sets are consistent, p5-sy_full leads to generate the closest F0 values to the

analysis-by-synthesis ones.

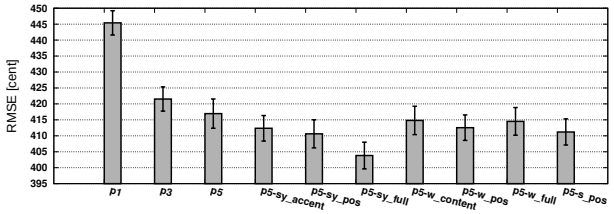


Figure 4: Global RMS error for the F0 component

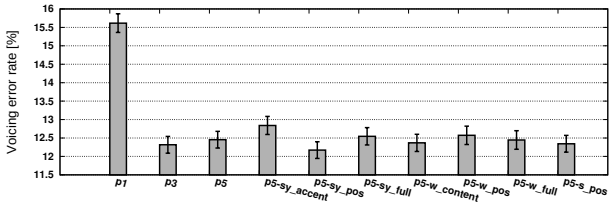


Figure 5: Global voicing error rate

In order to complete this analysis, we compute the RMS error in cent and the voicing error rate for each category of phones. The results are, respectively, presented in figures 6 and 7. In both cases two trends stand out. By comparing the contextual factor sets, we find that the improvement achieved by p3 relatively to p1 can be explained by specific categories like the voiced plosive (RMSE) or the voiced liquid (voicing error rate). Considering the diphthong, we cannot conclude as the number of frames is low comparing to other phonetic categories.

As we just mentioned, differences in the quality of the F0 modeling between phonetic categories can be observed. Unvoiced plosive and unvoiced fricative modelings are clearly worse than the others. This statement is valid in both measures. However, the voicing error rates associated with voiced segments are below 5%. Generally, the boundary of unvoiced labeled segment corresponds to a voicing boundary. Using MSD [12] implies that, during the training stage, one frame contributes to the voiced or the unvoiced distribution. This leads to a strict voiced/unvoiced split which implies problems at voicing boundaries. These results confirm that problem and indicate that no contextual factor set cannot avoid it even if using some specific factors could reduce this problem.

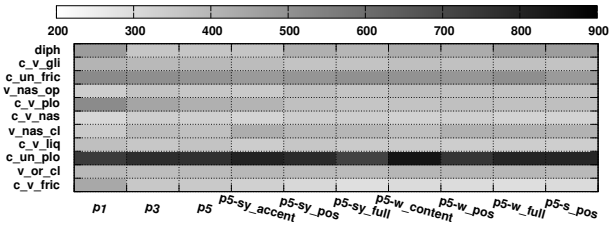


Figure 6: RMS error results by phoneme categories

3.3.3. Duration evaluation results

By applying the pairwise evaluation on the duration, we achieve results presented in figure 8. By comparing RMS error according to the contextual factor sets, the only significant difference

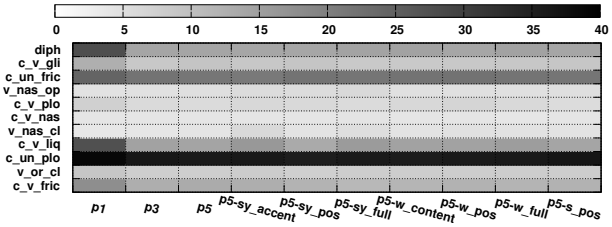


Figure 7: Voicing error results by phoneme categories

is provided by the p5 set in comparison of the p1 set. By comparison with GMM based evaluation, p3 results are more intermediate than the pairwise evaluation. These results indicate that the produced duration using p1 is not so distant than the ones provided using other descriptive features.

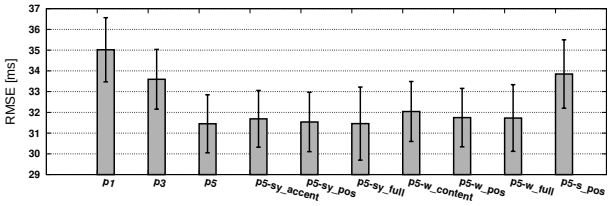


Figure 8: Global duration RMS error results

Finally, we focus the analysis by computing the RMSE for each phoneme. Results are presented in figure 9. The modeling of the phone /oe/ duration seems to be worse than other phones. However the confidence intervals, not detailed here, show that the difference is not statistically significant.

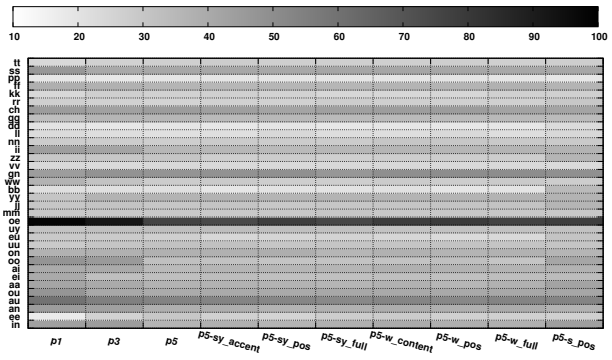


Figure 9: Duration RMS error results by phoneme

3.4. Objective evaluation conclusions

According to all results achieved by objective evaluations, the more suitable contextual factor set is p5-sy_full. For all evaluations, a clear improvement is introduced by taking into account the direct phonetic context (p3 relatively to p1). However, although phonetic features suffice to achieve a fine prediction of the segmental part (the best set is p3) and the duration (the best set is p5), pairwise evaluation indicates that the modeling of the fundamental frequency requires more contextual factors. These results are consistent with studies achieved for other languages like the one given in [9]. Furthermore, our results also indicate that differences in the modeling quality exist between phonetic

categories. This statement is obvious in case of the F0 modeling. Actually, based on the study presented in [13], we assume that these differences could be due to the use of the MSD in the standard version of HTS and do not depend on a contextual factor set.

4. Subjective evaluation

4.1. Evaluation procedure

A global subjective evaluation was conducted in order to complete the analysis of the objective evaluation results.

In this evaluation, seven signal sets are defined : the natural signal, the analysis-by-synthesis signal and the signals produced by HTS according to five contextual factor sets. The three first sets are p1, p3 and p5. They are used to assess the impact of the phonetic context in the synthesis. The last two sets are p5-sy_full, which objective evaluations tend to indicate that it is the more suitable, and p5-s_pos which is the most complete contextual factor set. Each signal set contains thirty utterances extracted from the test corpus and the average duration of each signal is about six seconds.

The goal of this test is to evaluate the overall quality of the synthesis using a MOS score. Nine listeners, working in speech processing, have done this test. One hundred stimuli have been presented to each listener. So, the evaluation test for each listener has been about thirty minutes.

4.2. Subjective evaluation results

The subjective evaluation results are detailed in figure 10.

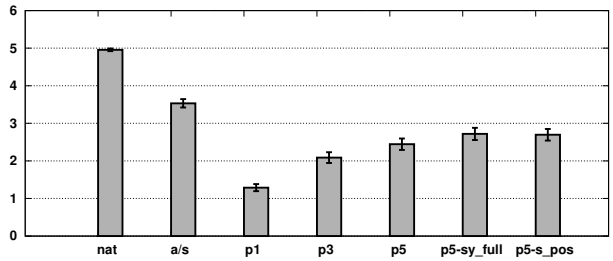


Figure 10: Results of the MOS evaluation

By comparing analysis-by-synthesis score against the score of the natural signal, we can notice that the listeners have perceived a signal damaging due to the signal parametrization. Among HTS synthesized signals, we can distinguish three sets: the signal synthesized using the combination p1, which has the lowest score; the signal synthesized using p3 whose quality is significantly improved against the signal p1 and the signals of the last three contextual factor combinations which have the highest score. However, signal deterioration due to the modeling is perceived, since all HTS synthesized signals are considered lower quality than the analysis-by-synthesis one.

As a significant improvement on the signal is perceived between p3 and p1, we assume that the modeling of each acoustic parameter is improved by taking into account the direct phonetic context. As for p5 and p5-sy_full, we assume that a better quality of the fundamental frequency modeling, done by HTS, is achieved by using syllable informations in the contextual factor sets. However, the subjective evaluation also confirms the results of the objective evaluations since no improvement was perceived between p5-sy_full and p5-s_pos.

5. Conclusion

In this paper, we have proposed an experimental protocol to objectively evaluate the synthesis achieved by HTS. This protocol is based on two complementary methods. The first one uses a GMM to model generated coefficient space and enables to assess the likelihood of the reference data according to this space. The second method relies on pairwise distances in order to carry out a more detailed analysis of the modeling achieved by HTS.

Using this protocol, we analyzed the closeness between the coefficients generated by HTS and those stemming from the STRAIGHT analysis for French synthesis. Experimental results suggest that using other descriptors than the phonetic and syllable context may be useless to improve the modeling achieved by HTS for this corpus. Based on the current methodology, further work must be achieved to analyze deeply the modeling achieved by HTS.

6. References

[1] H. Zen and T. Toda, “An overview of Nitech HMM-based speech synthesis system for blizzard challenge 2005”, Eurospeech, pp1957-1960, 2005.

[2] T. Fukada, K. Tokuda, T. Kobayashi and S. Imai,“An adaptive algorithm for mel-cepstral analysis of speech”, ICASSP, pp137-140, 1992.

[3] H. Kawahara, I.Masuda-katsuse and A. De Cheveigné, “Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds”, Speech Communication, 27, pp187-207, 1999.

[4] Tokuda, K., Zen, H. and Black, A. W., “An HMM-based speech synthesis system applied to english”, ICASSP, pp227-230, 2002.

[5] King, S. and Karaiskos, V., “The Blizzard challenge 2010”.

[6] Chen, Y.-n., Yan, Z.-j. and Soong, F. K., “A perceptual study of acceleration parameters in HMM-based TTS”, Interspeech, 2010.

[7] Silén, H., Helander, E., Nurminen, J. and Gabbouj, M., “Analysis of duration prediction accuracy in HMM-based speech synthesis”, Speech Prosody, 2010.

[8] Watts O., Yamagishi, J. and King, S., “The role of higher-level linguistic features in HMM-based speech synthesis”, Interspeech, 2010.

[9] Yokomizo, S., Nose, T. and Kobayashi, T., “Evaluation of prosodic contextual factors for HMM-based speech synthesis”, Interspeech, pp430-433, 2010.

[10] Astrinaki, M., d’Alessandro, N., Picart, B., Drugman, T. and Dutoit, T., “Reactive and continuous control of HMM-based speech synthesis”, SLT, pp252-257, 2012.

[11] Boeffard, O., Charonnat, L., Le Maguer, S., Lolive, D. and Vidal, G., “Towards fully automatic annotation of audiobooks for TTS”, LREC, 2012.

[12] Tokuda, K., Masuko, T., Miyazaki, N. and Kobayashi, T., “Hidden Markov models based on multi-space probability distribution for pitch pattern modeling”, ICASSP, 1999.

[13] Yu, K., Toda, T., Gasic, M., Keizer, S., Mairesse, F., Thomson, B. and Young, S., “Probablistic modelling of f0 in unvoiced regions in hmm based speech synthesis”, ICASSP, 2009.