INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

On the suitability of vocalic sandwiches in a corpus-based TTS engine

David Guennec, Damien Lolive

IRISA - University of Rennes 1 - Lannion, France
{david.guennec, damien.lolive}@irisa.fr

Abstract

Unit selection speech synthesis systems generally rely on target and concatenation costs for selecting the best unit sequence. The role of the concatenation cost is to insure that joining two voice segments will not cause any acoustic artefact to appear. For this task, acoustic distances (MFCC, F_0) are typically used but in many cases, this is not enough to prevent concatenation artefacts. Among other strategies, the improvement of corpus covering by favoring units that naturally support well the joining process (vocalic sandwiches) seems to be effective on TTS. In this paper, we investigate if vocalic sandwiches can be used directly in the unit selection engine when the corpus was not created using that principle. First, the sandwich approach is directly transposed in the unit selection engine with a penalty that greatly favors concatenation on sandwich boundaries. Second, a derived fuzzy version is proposed to relax the penalty based on the concatenation cost, with respect to the cost distribution. We show that the sandwich approach, very efficient at the corpus creation step, seems to be inefficient when directly transposed in the unit selection engine. However, we observe that the fuzzy approach enhances synthesis quality, especially on sentences with high concatenation costs.

Index Terms: concatenation cost, corpus-based TTS, unit selection

1. Introduction

In recent years, research in text-to-speech synthesis essentially focused on two major techniques. The statistical parametric approach (SPSS), which mainly includes HMM and DNN-based systems [1, 2, 3], is the most recent and has been the focus of many academic work in recent years. This method offers advanced control on the signal and produces very intelligible speech but generated voice lacks naturalness. The historical one, unit selection, is a refinement of concatenative synthesis [4, 5, 6, 7, 8, 9]. Sound created with this method features high naturalness and its prosodic quality is unmatched by other methods, as it basically concatenates speech actually produced by a human being. While most industrial TTS systems rely on unit selection, this method has its drawbacks, for instance the difficulty to force prosody and the possibility to get concatenation artefacts penalizing intelligibility.

In the formulation of the unit selection problem, a unit is a list of contiguous spectral segments (in the speech corpus) fitting a portion of the target sequence of phonemes. In order to discriminate the segments coming from the corpus that fit the requirements expressed via the target sequence, the usual method [5] is to rank the units by evaluating the context matching degree (target cost) and the risk of creating an artefact if concatenating the unit (concatenation cost) via balanced cost functions. The concatenation cost typically relies mainly on acoustic features (MFCC, F_0) [10, 11] to evaluate the level of spectral resemblance between two voice stimuli on and around the concatenation point. As for now, concatenation costs are far from being perfect and audible artefacts appear both in commercial and research TTS systems, even after post-concatenation processing. A few analyses, for example [12], showed that these artefacts occur more often on some phoneme than others. For instance, phonemes with high context-dependency (*e.g.* liquids) might show substantial inter-occurrence spectral variability [13], which is particularly dangerous for unit selection, especially because joining is usually done on phone centers (*i.e.* diphone boundaries). This being considered, some authors tried to use phonologically motivated rules to prevent joining on "risky" phonemes. For instance, in [12] the authors successfully tested a penalty system based on the phonological class of candidates to concatenation. A refined version of this idea was used by D. Cadic in the context of recording-script construction in [14] to favor covering of what has been called "vocalic sandwiches", also with success.

In this article, we study the impact of vocalic sandwiches back in the concatenation cost of a modern unit selection system, when a corpus was not created using the "sandwich" process described in [15]. We use the 3 phonologically-based phoneme clusters defined by [14] to forbid concatenations on phones believed to often cause joining artefacts. Believing this direct transposition marginalizes acoustic concatenation costs, we develop an enhanced version that softens penalties. This is done through the use of a fuzzy function that relaxes the penalty based on the acoustic concatenation cost distribution. It allows to smoothen the constraints imposed by sandwich penalties.

The main impact of this study is to improve TTS in the case of less controlled data, such as audiobooks, by transposing a constraint originally proposed for the corpus creation step directly into the TTS engine. The challenge is to know if the efficiency obtained at corpus building level can be found also at unit selection level. To that respect, unit selection makes it much simpler than SPSS to add the sandwich feature and test its efficiency. Experiments show the efficiency of the proposed approach and its suitability for corpus-based approaches at a low cost.

In this paper, the study is on French language but our conclusions should apply to other languages.

The remainder of the paper is organized as follows. In section 2, we briefly present the concept of vocalic sandwich. Section 3 first presents the TTS system on which our experiments are made. Then the integration of sandwiches into the system is discussed. Finally, we describe a fuzzy enhancement of the sandwich system. In section 4, we first describe our test data and then our experimental protocol. The experiments and their results are then presented. They are discussed in section 5.

http://dx.doi.org/10.21437/Interspeech.2016-1222

2. Enhancing speech corpora with vocalic sandwiches

Analysis of sentences containing artefacts shows that concatenation on some phonemes, especially vowels and semi-vowels, is more likely to engender artefacts than others (plosives and fricatives for example, especially unvoiced ones) [12]. Phonemes featuring voicing, high acoustic energy or important context dependency are generally subject to more distortions. Based on this claim, [14, 16] proposed a corpus covering criterion where the objective is to get a maximum covering of "sandwich units". A sandwich unit is a sequence of phonemes where one or several syllabic nuclei are surrounded by two phonemes considered as not likely to cause artefacts (we call it "resistant" to concatenation artefacts). A sandwich can therefore be formally defined as:

$$R(A^*VA^*)^+R \tag{1}$$

where + means 1 or more occurrences, * means 0 or more occurrences and R, A and V are the three following phonetic clusters, which Cadic *et al.* justifies in [14]:

V (vowel) : Vowels, on which concatenation is hardly acceptable.

A (acceptable) : Semi-vowels, liquids, nasals, voiced fricatives and schwa. These units are viewed as acceptable concatenation points, but still precarious.

R (resistant) : the remaining phonemes (unvoiced consonants, voiced plosives), where concatenation is definitely possible. The word "Resistant" is used in the following to describe units of this class.

3. Sandwiches in a unit selection engine

In this section, we describe how we integrate sandwich clusters into the unit selection concatenation cost, first with simple penalties and then with a much more refined fuzzy version. First though, we briefly present the TTS system used to implement the new costs. For a detailed description, refer to [17, 18, 19].

3.1. The IRISA TTS System

The IRISA TTS system, used for the experiments presented in this paper, relies on a unit selection approach realized with an optimal graph-search algorithm (here A* algorithm).

The concatenation cost between two units u and v is composed of MFCC (excluding Δ and $\Delta\Delta$ coefficients), amplitude and F_0 Euclidean distances:

$$C_c(u,v) = C_{mfcc}(u,v) + C_{amp}(u,v) + C_{F_0}(u,v), \tag{2}$$

where $C_{mfcc}(u,v)$, $C_{amp}(u,v)$ and $C_{F_0}(u,v)$ are the three sub-costs for MFCC, amplitude and F_0.

A set of preselection filters is used as binary target cost functions to filter candidate units from the corpus, including in the selection graph only those matching a set of linguistic/phonetic features. They rely on the assumption that if a unit doesn't respect the required set of features, it can't be used for selection. This means we have an absolute vision of what features units must match. One might argue this is not optimal, but by experience, more refined tuning doesn't prove to be better. The set of filters we use in this work is the same as in [18].

3.2. Phonologically motivated penalty based on sandwich classes

For the purpose of our study, we defined two penalization methods based on the three phonetic clusters defined in section 2.

We chose these clusters specifically because they are the same as those presented and justified in [14], though the choice of elements put inside each cluster is arguable, for example the choice of considering all vowels dangerous areas for joining.

As said earlier, using the phonetic class to constrain or penalize phonemes considered as problematic for concatenation is not a novel idea, and a few works can be cited, for example [20, 12]. However, in these works, costs and penalties are very constraining: either a unit respects the constraints set by the filters, or it dones not and thus gets drastically penalized (substentially reducing the probability that this unit is selected, regardless of its performance with the concatenation cost).

A key point of the idea we investigate here is that, because we do not want to add too many constraints in the cost function, we only defined 3 subsets of phonemes. The purpose of the penalty is not to act as a standalone cost, but simply to introduce knowledge that is not captured by the concatenation cost and then help achieve a finer ranking of units. Moreover, the proposed classes are based simply on basic linguistic/phonological knowledge and it may be necessary to adapt them depending on the language.

The first method for applying the penalty, called *sand*, is to give a fixed penalty $p(v)$ to each phoneme class: 0 for phonemes in R, a penalty slightly higher than the highest value of C_c observed in the corpus for all phonemes in A. Vowels (V) are given a huge penalty, big enough to prevent compensation by other costs in the candidate sequence. It corresponds to a penalization of candidate units based on the phonemes on which concatenation may be performed if choosing this unit. In this case, a new concatenation cost function C'_c is formulated as:

$$C'_c(u,v) = C_c(u,v) + K(u,v) \tag{3}$$

where $K(u,v) = p(v)$ is the penalty depending on the phoneme that begins the unit v as described before, which is the same as the phoneme ending u as we perform joining on diphone boundaries.

3.3. Fuzzy version

The second method, called *fuzzy-sand*, is to relax the penalty in certain cases. Thus, we introduce a fuzzy weighting function giving to each penalty a weight ranging between 0 and 1 as shown on figure 1. It describes how satisfying the candidate unit is with respect to its concatenation quality. Assuming MFCC, Amplitude and F_0 cost distributions follow normal distributions, we define two thresholds for each sub-cost. For instance, the two thresholds $T^1_{F_0}$ and $T^2_{F_0}$ for the F_0 sub-cost may be defined as:

$$T^1_{F_0} = \mu_{C_{F_0}} - \sigma_{C_{F_0}} \tag{4}$$

$$T^2_{F_0} = \mu_{C_{F_0}} + \sigma_{C_{F_0}} \tag{5}$$

Formally, the fuzzy function is defined, for the F_0 sub-cost:

$$f_{F_0}(u,v) = \begin{cases} 0 & \text{if } C_{F_0}(u,v) < T^1_{F_0}, \\ 1 & \text{if } C_{F_0}(u,v) > T^2_{F_0}, \\ 1.0 - \frac{(T^2_{F_0} - C_{F_0}(u,v))}{(T^2_{F_0} - T^1_{F_0})} & \text{otherwise.} \end{cases} \tag{6}$$

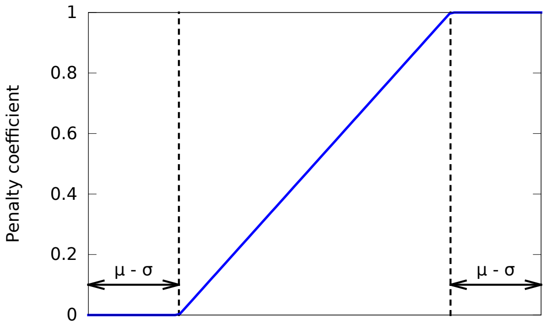

Figure 1: Fuzzy function over the distribution of sub-costs. The weight 0 (resp. 1) is given to units that have a concatenation costs approximately among the 15% lowest (resp. highest) costs. Between these thresholds, the weight increases linearly.

The choice for that tolerance interval is motivated by the observation of real cost distributions. To be complete, the choice of the thresholds should be differentiated depending on the type of sub-cost and optimized separately. Finally, the penalty is modified in the following way:

$$K(u,v) = (f_{mfcc}(u,v)+f_{amp}(u,v)+f_{F_0}(u,v))*p(v) \quad (7)$$

where $f_{mfcc}(u,v)$, $f_{amp}(u,v)$ and $f_{F_0}(u,v)$ correspond to the fuzzy functions of the form described in figure 1 respectively for MFCC, amplitude and F_0. With those functions, the main idea is to decrease the penalty when the unit has a concatenation sub-cost value which is statistically among the best ones. These distributions are estimated using the voice corpus by computing concatenation sub-costs for F_0, amplitude and MFCC using all units present in the corpus.

To sum up, if concatenation cost is above the higher threshold then we definitely have to apply the full penalty as the unit considered is among worst possible units. Between the two thresholds, we augment progressively the penalty as the concatenation cost increases.

4. Experimental evaluation

In this section, we first present our test data, then our experimental protocol and finally the results of our experiments. Results concerning *sand* and *fuzzy-sand* are put in distinct sections. For the experiments, we use a system called *baseline*, identical with *sand* and *fuzzy-sand* but without the sandwich feature.

4.1. Test data

For test purposes, we used two voices. The first one, called *Audiobook*, was built from an expressive audiobook. The speaker is a male with a low pitch (average F_0 is 87Hz). It contains 9h59' of speech and is sampled at 44.1kHz with lossless encoding. The voice was automatically annotated using the process described in [21] and using the ROOTS toolkit [22]. It is composed of 3139 utterances, with 353691 phones and 22727 non speech sounds. The second corpus, named *IVS*, was recorded for TTS purposes within an Interactive Vocal System with a hand-made recording script. The script aims at covering all diphonemes present in French and comprises the most used words in telecommunications vocabulary. The mean F_0 for the Female speaker is 163Hz. The corpus is composed of 7655 utterances, 238820 phonemes and 20407 non speech sounds for 7h06' speech. The recording is sampled at 16kHz (lossless encoding, 1 channel).

The evaluation corpus is a set of 27141 French sentences extracted from a wide variety of audiobooks, featuring many different styles, the same we used in [23].

4.2. Evaluation process

We synthesized our test set of 27141 sentences for our 3 systems (*baseline*, *sand* and *fuzzy-sand*). In order to evaluate the two sandwich concatenation cost adaptations presented earlier, we carried out a total of 12 AB listening tests split in 3 groups of four tests:

Random sentences : 4 tests where the sentences are picked up randomly among those generated in our test set. This serves as a baseline evaluation which aims at studying if the sandwich systems are, in average (*i.e.* in general), an enhancement over *baseline*.

Most different sentences : 4 tests where the most different synthesized stimuli pairs are chosen. Choice of the most different stimuli is made using DTW, as we presented in [23]. It aims at revealing differences that might have been obscured by the first set of tests by comparing the stimuli that are the most impacted by sandwich methods. If tested methods are worse than *baseline* in these tests, this methodology allows us to say sandwiches have mostly a negative impact on TTS, or the reverse if results are in favor of sandwiches.

Sentences with highest concatenation cost : 4 tests where the sentences are the ones that feature the biggest concatenation costs for the *baseline* system. They correspond to the sentences that most need improvement, and thus the primarily target we wish to enhance with the sandwich costs. If these sentences are not enhanced, this most likely means that sandwiches are inefficient as their purpose is to prevent disastrous concatenations more than enhancing joining quality.

Each test was made by 10 expert testers, each one evaluating 10 distinct stimuli pairs. 100 stimuli pairs are evaluated in total (all 100 synthesized from distinct sentences), each tester evaluating his own set of stimuli. In every test, the same question is asked: "Which of the two sample offers, for you, the best global quality?". The testers have to choose between the three following answers: A, B or Indifferent. For the last set of tests (high concatenation costs), a second question is asked at the same time, this time over concatenation quality. We made the choice of expert testers in particular because of this last question. For each set of 4 tests, we carry out two tests using *IVS* voice and two with *Audiobook*. For each voice, one test compares system *baseline* with *sand*, the other with *fuzzy-sand*. Test conditions are studio-like and follow ITU-T recommendations.

4.3. Results

Table 1 presents the results of the tests for the comparison *baseline versus sand*. Each line corresponds to one AB test. Column 1 indicates the voice used for the test and column 2 the selection method for the test sentences ("R." for random, "DTW" for most different and "C. C." for highest concatenation cost). Column 3 refers to the question asked during the test: either "C. Q." for the question on concatenation quality or "G. Q." for the assessment of global quality. Using the same representation, results for the *fuzzy-sand* method are presented on table 2.

Table 1: Results for the AB listening tests for the *sand* system. Lines concerning tests on random sentences have the mention "R." in the second column. "DTW" is for tests with most different sentences and "C. C." for tests on sentences of *baseline* with the highest concatenation costs. Column 3 displays "G. Q." when the question was on global quality and "C. Q." when it was on concatenation quality only.

			Answers		
			Base	*sand*	Indifferent
IVS	R.	G. Q.	**45%**	34%	21%
	DTW	G. Q.	31%	34%	**35%**
	C. C.	C. Q.	33%	30%	**37%**
		G. Q.	30%	**35%**	**35%**
Audiobook	R.	G. Q.	38%	**39%**	23%
	DTW	G. Q.	**47%**	32%	21%
	C. C.	C. Q.	**38%**	31%	31%
		G. Q.	**39%**	30%	31%

Table 2: Results for the AB listening tests for the *fuzzy-sand* system. Please refer to table 1 caption for explanation of the table.

			Answers		
			Base	*fuzzy.*	Indifferent
IVS	R.	G. Q.	35%	**40%**	25%
	DTW	G. Q.	31%	**48%**	21%
	C. C.	C. Q.	20%	**59%**	21%
		G. Q.	27%	**49%**	24%
Audiobook	R.	G. Q.	**43%**	42%	15%
	DTW	G. Q.	42%	**46%**	12%
	C. C.	C. Q.	33%	**38%**	29%
		G. Q.	36%	**43%**	21%

5. Discussion

First, if we compare the behavior of the systems regarding the number of concatenations[1], we find that *fuzzy-sand* lead to a larger number of concatenations (2264 concatenations with *IVS* and 2046 with *Audiobook*) than both *baseline* (1831 and 1663) and *sand* (1838 and 1663). This is not a problem: two well made (and therefore inaudible) concatenations are worth much more than one failed joining. When looking at the phoneme classes on which concatenations are made, we see that *sand* and *fuzzy-sand* produce much more concatenations on robust (cluster R) phonemes: *baseline* leads to 582 concatenations on phonemes from R with *IVS* against 1025 for *sand* and 1095 for *fuzzy-sand*. With *Audiobook*, the same numbers are: 606 for *baseline*, 865 for *sand* and 907 for *fuzzy-sand*. This is the proof the two methods work as expected. In addition, *fuzzy-sand* also causes substantially more concatenation on A and V clusters, as intended.

Second, from the listening tests results, we observe that the *sand* approach seems largely inefficient (at best) when integrated directly in the unit selection engine, as shown in every test made with the method. In some cases, it was even counterproductive: the AB test on random sentences for *IVS* clearly show it, and the same conclusion can be observed on 3 tests out of 4 concerning *Audiobook* voice. What is also noticeable is the high quantity of "indifferent" ratings, proof that the difference between the two systems isn't very clear. So we can say that, if sandwiches proved useful for the construction of a recording script (*cf.* [14]), they prove inefficient, or even counterproductive when directly integrated into the concatenation cost.

On the contrary, for almost every test with *IVS* voice, a clear superiority of the *fuzzy-sand* approach can be observed. The result is also observable for *Audiobook* voice, though with a smaller gap. *Audiobook* voice faring better than *IVS*, the lesser difference for the first one seems logical. The explanation for *Audiobook* faring better than *IVS* is that: (1) it is expressive while *IVS* is neutral and (2) *Audiobook* audio files sampling frequency is higher (44,1 *vs.* 16kHz). Concatenation quality is perceived better with *fuzzy-sand*. The number of "Indifferent" answers is also consistently lower for *fuzzy-sand*, meaning that differences are more easily felt. In conclusion, *fuzzy-sand* approach proves to be effective thanks to the degree of flexibility it adds in regard to *sand* method. In particular, we observe that *fuzzy-sand* ranking is between *sand* and *baseline*, which means it alters *baseline* ranking, based solely on acoustic measures that we know are imperfect, but not as much as *sand* (which completely changes the ranking and loses the information of acoustic measures). It is also interesting to see that the controlled corpus *IVS*, was more affected by sandwiches than *Audiobook*, which is completely uncontrolled. The question this raises is the following: is it the quality of *Audiobook* voice or its uncontrolled nature that causes the observed lower performance of sandwiches for that voice?

We believe that the key to the success of all these measures (including *fuzzy-sand*), is a close integration in the concatenation cost. The penalty cannot hide the ranking provided by acoustic costs, and this for a good reason: these penalties aim at correcting acoustic rankings on key points, using expert knowledge (and it is exactly what sandwiches at corpus building level does). But a too constraining penalty system (*i.e. sand*), which is not a good concatenation cost on its own, causes a complete re-ranking of the system, hence the drop in quality.

6. Conclusion

In this paper, we presented a study of the impact of vocalic sandwiches back in the concatenation cost of a modern unit selection system, through two penalty-based systems. This penalty enables to avoid some artefacts during synthesis and its fuzzy version preserves the ranking made by acoustic components of the concatenation cost. The subjective experiments we conducted show a better performance for fuzzy version both for a neutral and an expressive voice. It shows that the concatenation cost does not capture all the perceptual information and that adding some preferences over the type of units to concatenate improves the synthesized speech quality. On the contrary, *sand* method, which fares well at script construction level, seems largely inefficient when integrated directly in the unit selection engine. Further improvement of the fuzzy method can be made. In particular, more advanced fuzzy patterns might be investigated. Further work should be conducted about the phoneme sets R, A and V. These subsets shouldn't be considered fixed and an investigation on how they compare with other classifications should be done. In particular, liquids and glides could be added to V as they are usually problematic. Investigating language dependence of those classes is another important task. Finally, it would be particularly interesting to activate the fuzzy penalty only when the concatenation cost magnitude gets considerable.

[1] All numbers are not provided here due to the lack of space.

7. References

[1] A. W. Black, H. Zen, and K. Tokuda, “Statistical Parametric Speech Synthesis,” *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07*, vol. 4, 2007.

[2] J. Yamagishi, Z. Ling, and S. King, “Robustness of HMM-based speech synthesis,” in *Science And Technology*, 2008, pp. 2–5.

[3] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “The Effect Of Neural Networks In Statistical Parametric Speech Synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Melbourne, 2015, pp. 4455–4459.

[4] Y. Sagisaka, “Speech synthesis by rule using an optimal selection of non-uniform synthesis units,” in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. Ieee, 1988, pp. 679–682.

[5] A. W. Black and P. Taylor, “CHATR: a generic speech synthesis system,” in *15th conference on Computational linguistics*. Association for Computational Linguistics, 1994, pp. 983–986.

[6] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. Ieee, 1996, pp. 373–376.

[7] P. Taylor, A. W. Black, and R. Caley, “The architecture of the Festival speech synthesis system,” in *Proc. of the ESCA Workshop in Speech Synthesis*, 1998, pp. 147—-151.

[8] A. Breen and P. Jackson, “Non-uniform unit selection and the similarity metric within BT’s Laureate TTS system,” in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[9] R. A. J. Clark, K. Richmond, and S. King, “Multisyn: Open-domain unit selection for the festival speech synthesis system,” *Speech Communication*, vol. 49, pp. 317–330, 2007.

[10] Y. Stylianou and A. K. Syrdal, “Perceptual and objective detection of discontinuities in concatenative speech synthesis,” in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 2001, pp. 837–840.

[11] D. Tihelka, J. Matoušek, and Z. Hanzlíček, “Modelling F0 Dynamics in Unit Selection Based Speech Synthesis,” *Text, Speech and Dialogue*, vol. 1, no. Springer, pp. 457–464, 2014.

[12] J. Yi, “Natural-sounding speech synthesis using variable-length units,” Massachusetts Institute of Technology, Tech. Rep., 1998.

[13] B. Lindblom, “Spectrographic study of vowel reduction,” *The Journal of the Acoustical Society of America*, vol. 35, no. November 1963, pp. 1773–1781, 1963.

[14] D. Cadic, C. Boidin, and C. D’Alessandro, “Vocalic sandwich, a unit designed for unit selection TTS,” in *Tenth Annual Conference of the International Speech Communication Association*, no. 1, 2009, pp. 2079–2082.

[15] D. Cadic, C. Boidin, and C. D’ Alessandro, “Towards optimal TTS corpora,” in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valetta, Malta*, 2010, pp. 99–104.

[16] D. Cadic and C. D’Alessandro, “High Quality TTS Voices Within One Day,” in *Seventh ISCA Workshop on Speech Synthesis*, 2010.

[17] D. Guennec and D. Lolive, “Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System,” in *17th International Conference on Text, Speech and Dialogue*, 2014, pp. 449–457.

[18] D. Guennec, J. Chevelu, and D. Lolive, “Defining a Global Adaptive Duration Target Cost for Unit Selection Speech Synthesis,” in *18th International Conference on Text, Speech and Dialogue*, Plzen, 2015, pp. 149—-157.

[19] P. Alain, J. Chevelu, D. Guennec, G. Lecorvé, and D. Lolive, “The IRISA Text-To-Speech System for the Blizzard Challenge 2015,” in *Blizzard Challenge workshop*, 2015.

[20] R. E. Donovan, “A new distance measure for costing spectral discontinuities in concatenative speech synthesizers,” in *ITRW*, 2001.

[21] O. Boeffard, L. Charonnat, S. Le Maguer, D. Lolive, and G. Vidal, “Towards Fully Automatic Annotation of Audio Books for TTS.” in *LREC*, 2012, pp. 975–980.

[22] J. Chevelu, G. Lecorvé, D. Lolive, G. L. Jonathan Chevelu, and D. Lolive, “ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections,” in *LREC*, 2014, pp. 619–626.

[23] J. Chevelu, D. Lolive, S. Le Maguer, and D. Guennec, “How to Compare TTS Systems: A New Subjective Evaluation Methodology Focused on Differences,” *Interspeech*, 2015.