Automatic Feature Selection for Acoustic-Visual Concatenative Speech Synthesis: Towards a Perceptual Objective Measure

Utpala Musti, Vincent Colotte, Slim Ouni, Caroline Lavecchia
Brigitte Wrobel-Dautcourt, Marie-Odile Berger

Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France
m.utpala@gmail.com {vincent.colotte,slim.ouni,caroline.lavecchia}@loria.fr
{brigitte.wrobel-dautcourt,marie-odile.berger}@loria.fr

Abstract

We present an iterative algorithm for automatic feature selection and weight tuning of target cost in the context of unit selection based audio-visual speech synthesis. We perform feature selection and weight tuning for a given unit-selection corpus to make the ranking given by the target cost function consistent with the ordering given by an objective dissimilarity measure. We explicitly perform feature elimination to reduce the redundancy and noise in target cost calculation based on an objective metric. Finding an objective metric highly correlated to perception should improve the quality of tuning. This is the purpose of the second part where we are making an attempt to such goal. Firstly, we present the human-centered evaluation done of the synthesized audio-visual speech and secondly, its preliminary analysis in relation to the objective evaluation metrics. This analysis of correlation between objective and subjective evaluation results shows interesting patterns which might help in designing better tuning metrics and objective evaluation techniques. The key point is to find a link between objective and perceptual measures.

Index Terms: Unit selection, audio-visual speech synthesis, target cost, target feature selection, weight tuning.

1. Introduction

Speech synthesis based on unit selection is widely used due to its capability of preserving the rich information in the speech signal. The key to the synthesis of 'natural' sounding speech is the selection of units which are perceptually suitable for the targets required for synthesis. For this selection of units, assigning a cost called the target cost to quantify the perceptual suitability is crucial [1]. This is important for the pre-selection of appropriate candidates from a typical corpus which generally has a large number of candidates. It is also necessary for the selection of the candidate sequence for the final synthesis. The formulation of the target cost function has been divided into Independent Feature Formulation (IFF) and Acoustic Space Formulation (ASF) [2]. In the former approach, only high level features which consist of linguistic and phonetic characteristics alone are used to describe speech units and the calculation of target cost is based on vectors expressing these assuming they are independent. The usage of high-level features allows the automatic selection of candidates with suitable prosodic characteristics rather than prediction based on prosodic models [3]. The latter approach i.e. ASF includes a representation of the required target in terms of speech parameters termed as partial-synthesis function [2].

Intuitively the best method for target weight tuning would be hand tuning as it directly addresses the perceptual performance objective. Several algorithms have been proposed which perform target cost weighting by hand tuning [4, 5, 6, 7]. The factors like inter-personal perceptual differences and time requirement for listening make it laborious. It is also constrained by its scope as the tuning can only be done on a small set of synthesized sentences. Thus, the global performance of a hand-tuned target cost is not guaranteed to be the best.

Automatic weight tuning of target cost overcomes the difficulties and disadvantages of hand tuning and performs comparable to the latter. Weight Space Search (WSS) as described in [1] is based on the comparison of the acoustic speech of the synthesized sentence and the natural utterance. Many approaches are based on the direct comparison of acoustic speech of real phonemic unit (diphone or phoneme) with a target description and those selected from the corpus [1, 8, 9]. Simultaneous target and join cost tuning is proposed in some of them [8, 9]. Each target feature accounts for the variations in the acoustic speech and their duration, based on this discriminative information, the features can be weighted [3]. Another approach to weight tuning is to view unit selection as a classification problem, in which instead of defining an objective function to account for the subjective speech quality, the classification error is taken as the objective function to be optimized [10].

For unit selection through IFF based target cost, having a mutually independent set of features is very important. Having a large set of features makes it practi-

cally impossible to cover all the possible feature combinations. In this paper we present a weight tuning algorithm which is applicable to any IFF based target cost function. It is an iterative algorithm for simultaneous feature elimination and weight tuning of the descriptive features retained. The algorithm though not specific to acoustic synthesis, is actually developed within a framework of Acoustic-Visual (AV) speech synthesis. A target cost function is evaluated based on the comparison of the candidate ordering by it and that based on an objective dissimilarity calculated using acoustic and visual features [8]. We perform explicit feature elimination which to the best of our knowledge has not been presented in any of the past literature. This is the first contribution of this paper. This is different to the approach where a target features take low or negligible weight in the target cost function, which is implicit feature elimination. The second contribution of this paper is an approach for extracting patterns of speech perception based on which users judge speech quality from any subjective evaluation results. It is illustrated through a preliminary analysis of subjective evaluation results done on the speech synthesized by our AV speech synthesis system. The ultimate goal could be to find an objective metric highly correlated to perception to improve the quality of tuning. This approach is simple when compared to designing specific experiments for this purpose [11] and can be useful in extracting global patterns. In section 2, we present our feature selection and weight tuning algorithm for target cost function. In section 3, we describe the execution of this algorithm for our system and give the summary of features selected for audio-visual speech synthesis in French. In the second part of the paper, we describe the human-centered evaluation of the synthesized audio-visual speech and in section 5, we then present a preliminary analysis of the subjective evaluation results in comparison with the objective evaluation results calculated through the comparison of synthesized and real speech signals.

2. Feature selection and weight tuning

A general target cost function $C(t_i, u_i)$ in a IFF paradigm is calculated as the weighted sum of the constituent feature costs $C_\rho(t_i, u_i)(\rho = 1, ..., F)$ by the comparison of the elements of the target t_i and candidate feature vectors, where F is the number of target features and w_ρ is the weight of feature ρ: $C(t_i, u_i) = \sum_{\rho=1}^{F} w_\rho C_\rho(t_i, u_i)$. Alternative formulations are also possible but we assume the above formulation of target cost function. This target cost function has a role not only in the initial preselection of units but also in the final selection from preselected candidates which is operated by resolving the lattice of possibilities using dynamic programming algorithm, where target cost is classically combined with join cost (acoustic and/or visual). The join cost function is important for the reduction of concatenation artifacts. The combination of these costs can lead to select a candidate with not necessary the best target cost if we favor the join cost. In this paper, we only focus on the target cost irrespective of joint cost.

The target feature vector is supposed to implicitly describe the speech realization of a target or a candidate. During synthesis, in a IFF paradigm, the target specification only has the target feature description but no acoustic or visual speech realization (see 3. for the list of used features). But, two speech units which are similar (feature vector) are perceptually suitable for each other and hence their target cost function should be low. We use this comparison in acoustic/visual domain to evaluate the target cost function. For this purpose we define the following two functions: (1) dissimilarity between two speech units, similar to [8], (2) disorder with respect to a target cost function for its evaluation. Then, we summarize our feature selection and target feature weight tuning algorithm which we refer to as selection-tuning algorithm.

2.1. Dissimilarity between two units

The dissimilarity measure $D(u, v)$ between units u and v of a particular phoneme p is defined as follows:

$$D(u,v) = \begin{array}{l} w_{dur}D^{dur}(u,v) + w_{ac}D^{ac}(u,v) + \\ w_{vs}D^{vs}(u,v) + w_{f0}D^{f0}(u,v) \end{array} \quad (1)$$

Where, D^{dur}, D^{ac}, D^{vs} and D^{f0} represent dissimilarities in terms of the duration, acoustic speech, visual speech and f0 of the units and w_{dur}, w_{ac}, w_{vs} and w_{f0} are the weights given to these respective components. The duration dissimilarity D^{dur} is calculated as the normalized difference between the durations of the two units. For the other three components (acoustic, visual and f0); the RMSE (root mean squared error) is calculated between two trajectories normalizing the length of the two trajectories by simple linear interpolation.

2.2. Disorder

Consider a unit t as the target, and two units u and v as candidates from the corpus whose dissimilarity measure with respect to t is $D(t, u)$ and $D(t, v)$. Then, for an ideal target cost the following condition should hold good :

$$D(t,u) \diamond D(t,v) \quad \Leftrightarrow \quad C(t,u) \diamond C(t,v) \quad (2)$$

where, $\diamond \in \{<, =, >\}$. Thus, the disorder δ with respect to the target t and the two candidates u and v is defined as follows:

$$\delta_t(u,v) = \begin{cases} 0 & \text{if (2) verified} \\ |D(t,u) - D(t,v)| & \text{else} \end{cases}$$

Let U_p be the complete set of candidates in the corpus for a given phonemic label p. Taking each of the units as a target at a time (see Table 1) and all the others as candidates, the *total disorder* for that phoneme is

Ranking		Disorder calculation		
Target Cost	Dissimilarity			
$C(t,c_1)$	$D(t,c_1)$	$c_1:$	$\delta_t(c_1,c_2)+\delta_t(c_1,c_3)=$	$0+0$
$C(t,c_2)$	$D(t,c_3)$	$c_2:$	$\delta_t(c_2,c_1)+\delta_t(c_2,c_3)=$	$0+\lvert D(t,c_2)-D(t,c_3)\rvert$
$C(t,c_3)$	$D(t,c_2)$	$c_3:$	$\delta_t(c_3,c_1)+\delta_t(c_3,c_2)=$	$0+\lvert D(t,c_3)-D(t,c_2)\rvert$

Table 1: Disorder calculation. From the corpus consider four units of the same phonetic label, a target t and three candidates $\{c_1,c_2,c_3\}$. $D(t,c_i)$ and $C(t,c_i)$ are the dissimilarity and the target cost between the target t and candidate c_i. For the given target, the dissimilarity based ordering and the target cost based ordering of candidates is compared to calculate the disorder. The total disorder with respect to target t is the sum of the fourth column.

calculated for a particular target cost as follows: $\Delta = \sum_t \sum_{(u,v)} \delta_t(u,v)$, where, $u,v,t \in U_p$ and $t \neq u \neq v$. By this way, The total disorder measures the difference between the ranking given by the target cost and given by the dissimilarity measure. Ideally, the target cost and the dissimilarity measure should follow the same ranking. If it is not the case that means there is a disorder. In the following sections, we refer to this *total disorder* as *disorder*.

2.3. Algorithm

The goal of the algorithm is to tune the target costs by synchronize the target cost with the dissimilarity measure (same ranking). The main idea of the algorithm is that each target feature has some contributing information which gets reflected in the speech realization implicitly. When a feature is removed from the target cost function, its selection accuracy will deteriorate, and the extent to which it deteriorates quantifies the feature's importance. The information contributed by a feature is measured by determining the difference in disorder when the feature was included and excluded from a target cost. A feature is considered to add information (resp. noise) if the disorder increases (resp. decreases) when excluded from the target cost function. Those features which add information, their weights increase proportional to their contribution. Features adding noise, their weights decrease until they become contributing features; if a feature adds only noise continuously, it is eliminated from the feature set after a fixed maximum number of iterations. The execution starts by assigning same weights to all the features and terminates when either there is no change in feature weights or after a fixed number of maximum iterations. Fig.1 shows the evolution of disorder and an example of weight evolution of one feature.

3. Application to AV target cost function tuning

We applied the feature selection and weight-tuning algorithm to our speech synthesis system (see [12, 13] for a detailed presentation). Our AV corpus used for executing this algorithm has a total of 319 sentences which represents a total of 14634 diphones and includes a good variety of the most frequent diphones. The visual speech is represented by 12 principal components. For acoustic

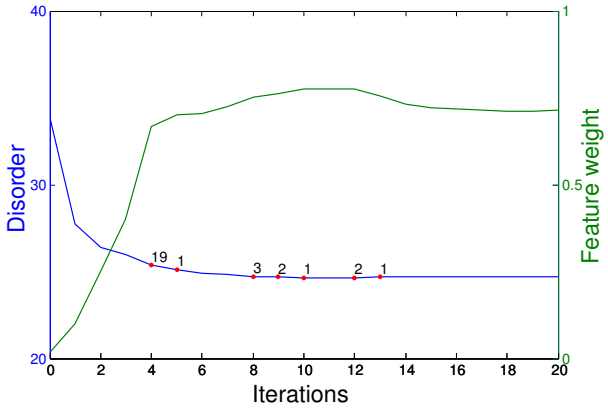

Figure 1: *Figure shows the change in total disorder and weight for the feature 'Syllable position in Rhythm Group' considering duration alone in the dissimilarity measure for the phoneme a. It shows the number of features eliminated at various iterations from an exhaustive set of 50 features.*

speech, we used 13 MFCC extracted every 10ms and f0 every 5ms.

The tuning has been done for five weight combinations in the dissimilarity measure (see Eq. 1). Four combinations had only one dissimilarity measure (duration, visual, acoustic and f0) i.e., only one of the weights $\{w_{dur}, w_{ac}, w_{vs}, w_{f0}\}$ being 1 and all others 0. The fifth combination was $(.25,.25,.25,.25)$ (i.e. equal weights to all of them). For each weight combination, the algorithm has been run separately for all the phonemes in the phoneme set using our corpus to obtain different target functions (different set of features and their weights) for different phonemes (see Fig. 1 for an example).

The initial exhaustive set of target features considered by us are the following:

- Phonetic context: previous and following phoneme's voicing, kind, manner of articulation, place of articulation and lip shape.
- Linguistic features of the current, previous and following phonemes' in terms of: kind of syllable; phoneme number in syllable; syllable position in word and rhythmic group (RG); syllable number in word, RG and sentence; word position in RG and sentence; word number in RG and sentence; RG position in sentence; position of nearest left and right silence.

Weights for vowels		
Feature	μ	σ
Place of articulation	0.36	0.18
Lip shape	0.14	0.19
Manner of articulation	0.09	0.09
Voicing	0.07	0.09
Kind	0.04	0.06

Table 2: Important phonetic features for visual speech (features with weights < 0.01 are not considered in these).

The above mentioned target feature set can be considered exhaustive for French language. We analyzed target features based on their relative importance for each of the constituent aspects included in the dissimilarity metric: pitch, local acoustic speech, duration and visual speech. We present for visual speech aspect ($w_{vs} = 1$ in the dissimilarity metric). Linguistic features can describe a current candidate or its left or right context. Phonetic features can describe a candidate left or right context. To analyze the results, we calculate the mean and the standard deviation of weights assigned to each feature by taking together the context and the current candidate. The weights are assigned such that the sum of the weights over all the target features is 1.

For visual speech, the total weight assigned to phonetic features is (on average) 0.69 for vowels and 0.88 for consonants (respectively, 0.31 and 0.12 for linguistic features). A summary of feature weights for visual speech is given in Table 2 for phonetic features for vowels. For vowels, place of articulation of the following and preceding phonemes are the most important features in the decreasing order of importance. The lip shape during articulation and manner of articulation of the contextual phonemes are also observed to be important. In the same way, consonants, lip shape of the following phoneme, lip shape of the preceding phoneme and place of articulation of the preceding phoneme are observed to be the 3 most important features in the decreasing order of importance. At linguistic level, syllable position in a word is an important feature for vowels. We have conducted the same study for the other constituents (Pitch, Duration, Local speech acoustics) [14].

The analysis of these selected features is in itself an interesting problem. The relative importance of the contextual features indicates that the right context is more important than the left. This is more pronounced in phonetic features weights. One of the possible interpretations of this is that the instances of anticipatory coarticulation is higher than the instances of carryover coarticulation in French. Word number in sentence has got eliminated for most of the phonemes as the corpus is not sufficient to establish any such relation. Numeric features in general have got lower weights which show that the relative position is more important than their exact position. The former features are size invariant. For example, 'syllable position in RG' does not depend on the total number of syllables in RG. But 'syllable number in RG' depends on the total number of syllables in RG. The selected features and their relative weights implicitly indicate the validity of the algorithm. For example, for pitch and duration, syllable position in RG, relative position of nearest left and right silence, syllable position in word are shown to be important. These features are known to be important for explaining many of the prosodic patterns in French.

With the fifth combination with equal weights to all the four constituents of the dissimilarity metric, the selected features contain the features which are important for all the four constituent aspects. We use these features and their weights determined in our synthesis system. We performed objective and human-centered evaluation for the synthesized audio-visual speech using these feature weights.

4. Evaluation

To evaluate our overall audio-visual speech synthesis system, word-level perceptual intelligibility and sentence-level subjective quality evaluation tests were conducted. 39 participants (15 females and 24 males) who are native French speakers between 19 to 65 years of age with normal auditory and visual abilities have participated in this experiment.

4.1. Intelligibility tests

Each human participant was presented with 50 one- or two-syllabic French words and asked to recognize and report the word. Some examples of the words that were presented include {anneau (ring), bien (good), chance (luck), pince (clip), laine (wool), cuisine(kitchen)}. Among these words, 11 were those which are present in the corpus (in-corpus words) which serve as a benchmark for the best-possible performance with the given corpus. These tests were done at two levels: (1) acoustic-only speech, (2) audio-visual speech. In each of these categories, the acoustic speech component was degraded to two noise levels. Hence, each word was played 4 times: (1) acoustic-only with low noise component (SNR of -6 dB), (2) acoustic-only with high noise component (SNR of -10 dB), (3) audio-visual with low noise (SNR -6dB), (4) audio-visual speech with high noise (SNR of -10 dB). The addition of noise also ensures that the listener pays attention to the visual modality of speech. The aim is to evaluate both visual and acoustic modalities, and also to estimate the advantage of audio-visual speech over acoustic-only speech. These noise thresholds were decided based on the several audio-visual perceptual experiments to strike a trade-off between these two objectives. The facial animation is shown as the 3D surface of the face using sparse mesh, which has the dynamics of facial deformations, but without the texture and color information [13]. Besides, the information regarding internal articulators, teeth and tongue is also missing from the ani-

	Audio		Audio-Visual	
	L.N.	H.N.	L.N.	H.N.
In-Corpus	0.69	0.59	0.72	0.65
Out-of-Corpus	0.40	0.34	0.45	0.40

Table 3: Mean intelligibility scores

	Q1	Q2	Q3	Q4	Q5
Overall	3.88	3.93	3.04	2.92	3.02
Out-of-Corpus	3.76	3.78	2.57	2.80	2.65
In-Corpus	4.80	4.91	4.56	3.67	4.32

Table 4: Mean MOS scores for the five questions

mations. Table 3 shows the mean intelligibility scores of in-corpus words and out-of-corpus words.

4.2. Subjective evaluation tests

Subjective tests were performed for the evaluation of the synthesis quality. 20 audio-visual sentences were played, out of which 7 sentences were real and the rest (13 sentences) were synthesized sentences which correspond to a subset of the test sentences we have for objective evaluation purpose. Just as in the case of intelligibility tests, the 7 real sentences serve as the best response that is possible with the corpus utilized for synthesis which affects various aspects of the synthesized speech like duration, phonetic coverage and facial speech rendering technique. For each of the stimulus, 5 questions were posed and participants were asked to give categorical responses based on the 5 point MOS scale (see Table 5). The first question (Q1) represents the synchrony in the acoustic and visual modalities. The second question (Q2) implicitly represents the prosody. Third and fourth questions (Q3 and Q4) are representative of the naturalness of acoustic and visual modalities respectively. The last question (Q4) is representative of the overall speech quality and pleasantness. The subjective evaluation results for in-corpus and out-of-corpus sentences are given in Table 4. The results to the question Q1 show that the audio-visual alignment is good, and the acoustic prosody is acceptable (Q2 results). It has to be highlighted that the prosody was generated without using any explicit model. The naturalness scores for voice seem to be low as shown in the Q3 results. These can be attributed to the relatively small size of the corpus and consequently the absence of some diphones in the corpus. On the contrary, the naturalness scores of facial animation (Q4 results) are high. This shows that articulatory dynamics are being represented well. Further, there might be a small component of the fact that the facial representation or 'human likeness' is not close to the uncanny valley and so participants are not very critical.

	Question	Categorical responses
Q1	Does the lip movement match the pronounced audio?	(5) Always – (1) Never
Q2	Is this sentence an affirmation (neutral reading)?	(5) Totally agree – (1) Not at all
Q3	Is the acoustic speech natural?	(5) Very natural – (1) Not natural
Q4	Is the facial animation natural?	(5) Very natural – (1) Not natural
Q5	Is the pronunciation of this sentence by the talking head pleasant?	(5) Very pleasant – (1) Not at all

Table 5: Five questions and the expected categorical responses.

5. Analysis of perceptual evaluation for better objective metrics

The per-sentence quality evaluation scores for each sentence were compared with the objective evaluation scored used during the system development incrementally. Correlation coefficient and Root Mean Squared Error (RMSE) calculated using PCA coefficients, MFCCs and F0 were used besides the segment duration ratios as the objective evaluation metrics. To investigate for the perceptually important segments which affect these subjective evaluation results, they were analyzed in comparison with the objective evaluation metrics. The analysis was based on the acoustic and visual modality. For this purpose different phoneme sets belonging to different categories were considered; like, all-phonemes, vowels, consonants, voiced phonemes, unvoiced phonemes, visible phonemes, visible vowels, not-visible phonemes etc. Visible phonemes are those which have identifiably unique visible articulation, like /p/, /o/ etc. The visible phoneme set includes those phonemes which are shown to have good recognition based on visual features. For test synthesized sentences; we have the real utterances, i.e. real acoustic and visual speech realization (these sentences are not uses in synthesis corpus). For these sentences, the objective evaluation metrics were calculated by comparing the synthesized and real utterances as follows. For each phoneme category, overall objective evaluation metrics mentioned were calculated. We refer to these metrics as consolidated metrics. As sometimes the subjective opinions can get affected by a few bad synthesis instances irrespective of a high overall performance, segment-wise objective evaluation metrics mentioned were also calculated and the minimum (undesirable) of each is determined. For example, if there are three vowels in a sentence, we keep the maximum of the RMSE as the representative of that sentence based on this metric (RMSE) and this phoneme category (vowel). We refer to these metrics as worst-case-based metrics.

With these objective metrics calculated, the subjective evaluation results for Q1 (AV synchrony), Q3 (acoustic speech naturalness), Q4 (visual naturalness) and Q5 (pleasantness of utterance) were correlated. This was an attempt to investigate the influential aspects which drive the perceptual opinion about the synthesized speech. The correlation results suggest the possibility of the following relations. A correlation between:

- Q1 scores (synchrony) and visible-vowels. This observation is based on Q1 scores and the consolidated correlation coefficients in visual and acoustic modality for visible-vowels.
- Q3 scores (acoustic speech naturalness) and worst-case f0 segments of voiced phonemes. This observation is based on the Q3 scores and worst-case-based f0 correlation. This is probably due to high sensitivity of human beings to prosody.
- Q3 scores (acoustic speech naturalness) and worst-case acoustic segments. This observation is based on the Q3 scores and worst-case-based acoustic speech correlation.
- Q4 scores (visual speech naturalness) and voiced phonemes and voiced-invisible phonemes. This observation is based on the Q4 scores and the consolidated visual speech correlations for voiced phonemes and voiced-invisible phonemes. This is probably due to human beings being critical towards coarticulation.
- Q5 scores (pleasantness of utterance) and f0. This observation is based on the Q5 scores and the worst-case-based f0 correlations for voiced phonemes. This is probably due to human beings being critical towards prosody.
- Q5 scores (pleasantness of utterance) and vowel and visible-vowel durations. This observation is based on the Q5 scores and the duration ratios for vowels and visible-vowel duration ratios. This is probably due to human beings being critical towards prosody.

This was a preliminary attempt to investigate for informative patterns. To draw definite conclusions, more rigorous and systematic analysis is necessary. Some expected patterns were not observed in these results. This might be due to the lack of test cases to bring out the relation between human perception and objective evaluation metrics.

6. Conclusion

In this paper, we presented a description of an algorithm for weight tuning and feature elimination of target cost function for unit selection based speech synthesis. The main goal was to keep independent and most informative features and tune their weights optimally. The method is generic: it can be applied for AV synthesis, which is our main focus, and also for acoustic-only synthesis. We presented a summary of the selected features for different aspects chosen to constitute the criteria for weight tuning. We described the objective and human-centered evaluation we performed on the synthesized speech. We described the further analysis we did to bring out correlation between the perceptual and objective evaluation results. We enlisted interesting correlation between objective and subjective evaluation results which if further explored can help in designing better objective evaluation metrics. The test sentences needs to be increased dramatically to bring out the correlation of human perception and different aspects of synthesized speech. This is being planned for future.

7. Acknowledgement

Our work was supported by the French National Research Agency (ANR - ViSAC (P.I. Slim Ouni) - Project N. ANR-08-JCJC-0080-01).

8. References

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, 1996.

[2] P. Taylor, *Text-to-Speech synthesis*. Cambridge University Press, 2009.

[3] V. Colotte and R. Beaufort, "Linguistic features weighting for a Text-To-Speech system without prosody model," in *INTERSPEECH*, 2005.

[4] G. Coorman, J. Fackrell, P. Rutten, , and B. V. Coile, "Segment selection in the L&H Realspeak laboratory TTS system," in *ICSLP*, 2000.

[5] F. Alías, L. Formiga, and X. Llorà, "Efficient and reliable perceptual weight tuning for unit-selection test-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept," *Speech Communication (Special issue on Perceptual and Statistical Audition)*, 2011.

[6] F. Alías, X. Llorà, I. I. Sanz, J. C. Socoró, X. Sevillano, and L. Formiga, "Perception-guided and phonetic clustering weight tuning based on diphone pairs for unit selection TTS," in *ICSLP*, 2004.

[7] H. Peng, Y. Zhao, and M. Chu, "Perceptually optimizing the cost function for unit selection in a TTS system with one single run of MOS evaluation," in *ICSLP*, 2002.

[8] L. Latacz, W. Mattheyses, and W. Verhelst, "Joint target and join cost weight training for unit selection synthesis," in *INTERSPEECH*, 2011.

[9] F. Alías and X. Llorà, "Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis," in *EUROSPEECH*, 2003.

[10] S. S. Park, C. K. Kim, and N. S. Kim, "Discriminative weight training for unit-selection based speech synthesis," in *EUROSPEECH*, 2003.

[11] C. Mayo, R. A. J. Clark, and S. King, "Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, 2011.

[12] A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger, "Setup for acoustic-visual speech synthesis by concatenating bimodal units," in *AVSP*, 2010.

[13] S. Ouni, V. Colotte, U. Musti, A. Toutios, B. Wrobel-Dautcourt, M.-O. Berger, and C. Lavecchia, "Acoustic-visual synthesis technique using bimodal unit-selection," *EURASIP J. Audio, Speech and Music Processing*, 2013, (in press).

[14] U. Musti, "Acoustic-visual speech synthesis by bimodal unit selection," Ph.D. dissertation, Université de Lorraine, 2013.