The IRISA Text-To-Speech System for the Blizzard Challenge 2017

Pierre Alain, Nelly Barbot, Jonathan Chevelu, Gwénolé Lecorvé,
Damien Lolive, Claude Simon, Marie Tahon

IRISA, University of Rennes 1 (ENSSAT), Lannion, France

pierre.alain@univ-rennes1.fr, nelly.barbot@irisa.fr, jonathan.chevelu@irisa.fr
gwenole.lecorve@irisa.fr, damien.lolive@irisa.fr
claude.simon.1@univ-rennes1.fr, marie.tahon@irisa.fr

Abstract

This paper describes the implementation of the IRISA unit selection-based TTS system for our participation to the Blizzard Challenge 2017. We describe the process followed to build the voice from given data and the architecture of our system. It uses a selection cost which integrates notably a DNN-based prosodic prediction and also a specific score to deal with narrative/direct speech parts. Unit selection is based on a Viterbi-based algorithm with preselection filters used to reduce the search space. A penalty is introduced in the concatenation cost to block some concatenations based on their phonological class. Moreover, a fuzzy function is used to relax this penalty based on the concatenation quality with respect to the cost distribution. Integrating a lot of constraints, this system achieves average results compared to others.

Index Terms: speech synthesis, unit selection

1. Introduction

In recent years, research in text-to-speech synthesis essentially focused on two major approaches. The first one is the parametric approach, for which HTS [1] and DNN-based systems [2] are now dominating the academic research. This method offers advanced control on the signal and produces very intelligible speech but with a low naturalness. The second approach, unit selection, is a refinement of concatenative synthesis [3, 4, 5, 6, 7, 8, 9]. Speech synthesized with this method features high naturalness and its signal quality is unmatched by other methods, as it basically concatenates speech actually produced by a human being.

The 2017 challenge is to build an expressive voice using children's audiobooks in English. The main difficulty with audiobooks, and in particular for children, is the change of characters and especially the imitation of animals (*i.e.* roars) as well as other sounds that may occur. For instance, in the data provided, a bell ringing signal is given to tell the child that he/she has to turn the page. Considering the expressivity of the voice, the different sounds and characters we can find in such books, the main challenges are phone segmentation and expressivity control.

In this paper we present the unit-selection based IRISA system for the Blizzard Challenge 2017. Basically, the system is based on preselection filters to reduce the acoustic unit space to explore and on a beam-search algorithm to find the best unit sequence. The objective function minimized by the algorithm is composed of a target cost and a join cost. The join cost relies mainly on acoustic features to evaluate the level of spectral resemblance between two voice stimuli, *on* and *around* the position of concatenation. For instance, distances based on MFCC coefficients and especially F0 are used [10, 11]. In particular, for the challenge, we have introduced a penalty on units whose concatenation is considered as risky. This follows the work of [12, 13] which showed that artefacts occur more often on some phonemes than others. For this purpose, we define a set of phoneme classes according to their "resistance" to concatenation. A phoneme is called resistant if the phones of its class are usually unlikely to produce artefacts when concatenated. This approach has been originally proposed in the context of recording script construction in [13] to favor the covering of what has been called "vocalic sandwiches".

Moreover, as audiobooks for children contain very expressive speech, one need a mean to control the degree of expressivity selected units have. To do so, we propose two things in our contribution. The first one is to introduce a prosodic model to predict how should be the prosody for each text segment. This is done using a DNN learned with the speaker's data. Predictions are then used in the target cost to rank units based on their prosodic properties. The second proposal is to build an expressivity score to evaluate how expressive a speech segment is in the acoustic space of the speaker. This score is then used to favor less expressive segments for narrative parts and more expressive segments during direct speech.

The remainder of the paper is organized as follows. Section 2 describes the voice creation process from the given data. Section 3 details the TTS system and further details are given in sections 4 and 5. Section 6 presents the evaluation and results.

2. General voice creation process

As in 2016, this year the challenge focuses on audiobook reading for children in English. The goal is then to build a voice based on approximately 6.4 hours of speech data provided as a set of wave files with the corresponding text. The recordings correspond to a set of 56 books targeted at children aged from 4 years old.

2.1. Data preparation and cleaning

The very first step has been to clean the text and make sure that it was corresponding to the speech uttered by the speaker. Moreover, all the quotation marks has been checked to insure an easy detection of boundaries between narrative and direct speech. Some parts corresponding to too expressive speech were discarded at this step to avoid later problems during synthesis. Despite of this, we still have preserved the major part of the expressive content. This work and the sentence level alignment has been done manually using Praat [14].

Finally, as the signals were provided using different formats, we have converted all the speech signals to standard WAV with a sampling frequency of 16kHz for further processing. F0 is extracted using the ESPS algorithm [15] while pitch marks

are computed using our own algorithm.

2.2. Segmentation and feature extraction

To build the voice, we first phonetized the text thanks to the grapheme-to-phoneme converter (G2P) included in *eSpeak* [16]. Then the speech signal has been segmented at the phone level using HTK [17] and standard forced-alignment. The acoustic models used for segmentation are learned using the data provided for the challenge.

Additional information is extracted from the corpus, like POS tags, syllables. Moreover, a label is associated to each word indicating if it is part of direct speech or not. The label is obtained based on the quotation marks in the text. The main idea with this label is to separate normal speech from highly expressive speech, usually present in dialogs.

Some prosodic features are also derived from the speech signal as energy, perceived pitch (in semi-tone) or speech rate. For those features, we compute minimum, maximum, average and standard deviation values at a word level. Those features are used during the synthesis process to better control the prosodic context associated to candidate segments.

All this information is stored in a coherent manner using the ROOTS toolkit [18]. All conversions and interactions between the different tools are also managed with this toolkit as, for instance, conversions from IPA (output of *eSpeak*) to the ARPABET phone set used in the synthesis engine.

3. The IRISA system

3.1. General architecture

The IRISA TTS system [19, 20], used for the experiments presented in this paper, relies on a unit selection approach with a beam-search algorithm. The optimization function is divided, as usually done, in two distinct parts; a target and a concatenation cost [4] as described below:

$$U^* = \underset{U}{argmin} \left(W_{tc} \sum_{n=1}^{card(U)} w_n C_t(u_n) + W_{cc} \sum_{n=2}^{card(U)} v_n C_c(u_{n-1}, u_n)\right) \quad (1)$$

where U^* is the best unit sequence according to the cost function and u_n the candidate unit trying to match the n^{th} target unit in the candidate sequence U. The search process is done using the beam-search algorithm using a beam of size 300. $C_t(u_n)$ is the target cost and $C_c(u_{n-1}, u_n)$ is the concatenation cost. W_{tc}, W_{cc}, w_n and v_n are weights for adjusting magnitude for the parameters. Sub-costs are weighted in order to compensate magnitudes of all sub-costs as in [21]. In practice, the weight for each sub-cost c is set to $1/\mu_c$, where μ_c is the mean sub-cost c for all units in the TTS corpus. The problem of tuning these weights is complex and no consensus on the method has emerged yet. [22] is a good review of the most common methods.

3.2. Join cost

The concatenation cost $C_c(u, v)$ between units u and v is composed of MFCCs (excluding Δ and $\Delta\Delta$ coefficients), amplitude, F0 and duration euclidean distances, as below:

$$C_c(u,v) = C_{mfcc}(u,v) + C_{amp}(u,v) + C_{F0}(u,v) + C_{dur}(u,v) + K(u,v),$$

Table 1: *List of features used in the target cost*

Phoneme position:	
LAST_OF_BREATHGROUP	
LAST_OF_WORD	LAST_OF_SENTENCE
FIRST_OF_WORD	
Phonological features:	
LONG	NASAL
LOW_STRESS	HIGH_STRESS
Syllable related features:	
SYLLABLE_RISING	SYLLABLE_FALLING

where $C_{mfcc}(u,v)$, $C_{amp}(u,v)$, $C_{F0}(u,v)$, $C_{dur}(u,v)$ are the sub-costs, resp., for MFCC, amplitude, F0 and phone duration. $K(u,v)$ is a penalty taking into account the estimated quality of the concatenation considering the distribution of the concatenation costs for phonemes of the same class. The computation of this penalty is detailed in [20, 23].

3.3. Target cost

For candidate units, we compute a numerical target cost built upon the following components:

- A linguistic cost computed as a weighted sum of the features given in table 1.
- A prosodic cost based on the euclidian distance between a set of basic prosodic features predicted by a DNN and the true value of candidate segments.
- An "expressivity" score used to control the level of expressivity of the candidates depending on their context. The underlying hypothesis is that we can rank the speech segments on an expressive scale and, for instance, favor candidates with high energy during direct speech while keeping more quiet candidates for narrative parts.

These three parts are summed up to result in the target cost. Finally, the weights W_{tc} and W_{cc} used in (1) to merge join and target costs are arbitrarily set.

In the following sections, we give more details on the two last sub-costs.

4. Prosodic target cost

In the case of story telling, the control of prosody is of first importance. Consequently, we tried to introduce a model learned on the speaker's data to predict some prosodic parameters for which we can compute a distance during the unit selection process. We chose to keep track of three discretized prosodic cues: speech rate (slow, normal, fast), F0 contour shape (rising, flat, falling), and energy level (low, normal, high). As an input to our model, we use 142 linguistic features such as the phone identity and some positional features (within the syllable, word, breath group and sentence).

The relationship between those input features and the output prosodic features is learned using a DNN. Based on empirical experiments, we decided to use a network with 3 hidden layers. The first one is a Bidirectional-LSTM layer with 256 units while the next two hidden layers are fully connected layers with 256 nodes each. The leaky rectified linear activation function is used for those layers. The network parameters are optimized using the RMSProp algorithm with a MSE loss function.

The coefficient of determination, or R^2, can be used to evaluate the ability of a model to predict real observed values. This score evaluates the proportion of output variance that is captured by the model. The possible values range from minus infinity to 1.0. A score of 0 means that the model outputs a constant value equal to the average of outputs. The best possible value is 1. In our case, the evaluation of the model gives R^2 scores of 0.95 on the training set, 0.92 on the validation set and 0.87 on the test set. Those results seem to show that the model is able to predict quite well the prosodic features.

During synthesis, the predicted values are used to evaluate the quality of candidate segments by computing an euclidian distance between predicted and real values. The resulting value is incorporated into the target cost as our prosodic cost.

5. Dealing with narrative/direct speech parts

Story telling, especially targeted at children, involves a lot of variations in expressivity. For instance, a great difference exists between narration and direct speech, i.e. when the character is speaking for himself like in a dialog. Changes can be made at same time on the timber and/or the prosody used by the reader to produce a living story and keep the attention of the listener.

5.1. Principle

To try to take into account such changes, we propose here to build a system enabling to give a "normality"/"expressivity" score to each word of the corpus used to build the TTS voice. The main idea behind this is (i) to characterize the normal way of speaking of the given speaker and, (ii) to give a score to each word based on its distance to normal events. In our case, the narrative sections, which represent the main part of the corpus, are considered as the normal way of speaking while direct speech parts are considered as outliers.

5.2. Expressivity score

To model this space of normal events, we use the energy (min, max, mean, std), perceived pitch (in semi-tone) and F0 (min, max, mean, std) features. One gaussian mixture model (GMM) is built per feature family using the scikit-learn toolkit [24]. The number of gaussian components per GMM is 8 at maximum and is controlled using BIC. We use a low number of gaussian components to avoid the specialization of some components for minor clusters that can be far from the majority classes. Other options might be chosen, such as the a posteriori elimination of gaussian components with a low weight (i.e. representing a low number of samples). As a consequence, common events should have a high likelihood for the model while words pronounced in a different way (e.g. with high energy or F0) should have a low likelihood.

The expressivity score S_{expr} is then computed as a linear combination of the probability of the word features w for each of the three models:

$$\begin{aligned} S_{expr}(w) &= -[\alpha_e logP(w|\mathcal{M}_e) + \alpha_t logP(w|\mathcal{M}_t) \\ &\quad +\alpha_f logP(w|\mathcal{M}_f)] \end{aligned}$$

where α_e, α_t and α_f are the mixing coefficients for energy, tone, rate and F0. $\mathcal{M}_e$, $\mathcal{M}_t$ and $\mathcal{M}_f$ are the corresponding GMM for each feature type.

The optimization of the mixing coefficients is done with a gradient descent on the narrative class only. Other kind of features have been tried, like speech rate but they were not relevant here.

5.3. Integration into the cost function

The next step is to compute the score for all the words in the corpus. During synthesis, two different target values are chosen for narrative and dialog parts. In the target cost, we add a subcost evaluating the distance between the target value and the true value for this score.

Ideally, we expect that a low target score will constrain the voice to remain less expressive while a higher target value will give the preference to more atypical segments.

One limit of this approach is that if the scores of two segments are low (resp. high), all we know is that these two segments are frequent (resp. infrequent) but we have no insight into the similarity of the two segments. Preliminary experiments have shown interesting results in some cases.

Another problem of this approach is that it can bring expressivity while introducing strong constraints on the selected segments. Depending on the content of the corpus, it can be harmful for the output quality. Notably, it can lead to inconsistencies during unit selection for instance concerning intonation or stress. This is what has been observed in the results for our system. In particular, the constraint is constant during the breath group while it could be better to adapt it in function of the corpus content and the choice of the other candidate segments in the sequence.

6. Evaluation and results

The evaluation assessed a number of criteria (overall impression, speech pauses, intonation, stress, emotion, pleasantness and listening effort) for book paragraphs as well as similarity to the original speaker, naturalness and intelligibility. The evaluation has been conducted for different groups of listeners: paid listeners, speech experts, and volunteers. In this section, we only give results including all participants. In every figure, results for all 17 systems are given. Among the system, we have the following : system A is natural speech, system B is the Festival benchmark system (standard unit selection), system C is the HTS benchmark and system D is a DNN benchmark. System Q is the system presented by IRISA.

6.1. Evaluation with paragraphs

Overall results are shown on figure 1 taking into account all listeners. For each criterion, our system achieves average results. These average results are likely to be explained by inconsistencies in the prosody and stress placement. A positive point is that the emotion criterion obtains a mark of 2.8 which seems to show that the proposed expressivity score has an impact.

6.2. Similarity to original speaker and naturalness

The similarity of the speech produced, as shown on figure 2, is among the average systems with a mean score of 2.8 and the median value at 3. Similarly, naturalness is also quite good as shown on figure 3 with an average of 3.1 and a median of 3. For naturalness, our system is comparable to the baseline festival system.

Despite of that, those results are far from the best systems. They seem to reinforce the conclusion that too many constraints have been introduced during the selection. Sometimes, the system performs very well but on average it makes many errors

Figure 1: *Mean Opinion Scores for the different criteria and the systems evaluated. "Natural" system is shown in yellow and IRISA system in red while other participants are in green.*

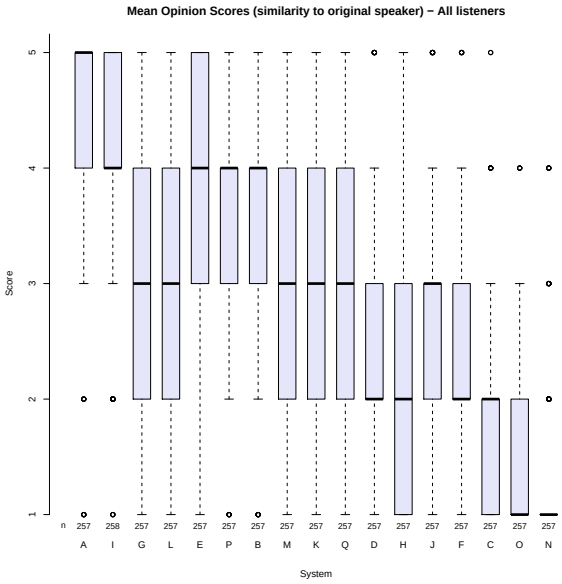

Figure 2: *Mean Opinion Scores, similarity to the original speaker evaluation, all listeners.*

penalizing the similarity and the naturalness criteria. Moreover, the downsampling to 16kHz of the speech signal may be a reason for the similarity degradation compared to our entry in 2016.

6.3. Intelligibility

Concerning intelligibility, our system is comparable to other systems with an average word error rate of 44%. Detailed results are given on figure 4. Compared to last year, the corrections we made have improved the intelligibility, even if our system is not performing well on other criteria.

7. Discussion

Despite of the improvements we added to our system, the results are not satisfying. After inspecting them and also the configuration of system, it appears that some elements can be corrected quite easily and seem to have a large impact on the synthesis quality. First, we have implemented a mechanism enabling to relax stress constraints in case not enough units are present in the right context. This mechanism introduces some inconsistencies in the stress placement as a lot of segments are not well represented in the corpus. By activating this threshold only when it is really needed (less than five units in the corpus), the stress placement seems to be improved, at least during informal listening tests.

Moreover, the expressivity score should be predicted word by word during synthesis instead of being chosen arbitrarily for an entire breath group. What appears here is that a constant target expressivity score may have an overall negative impact on intonation. In future work, we should focus on that particular problem. The introduction of a neural network to guide unit selection seems to work well and helped to control realized prosody thus avoiding very low score for intonation.

During the development of the expressivity score, we checked the ranking of the words informally by listening to the words with the highest and the lowest scores. Doing that helped us detect big segmentation errors and thus improve the quality of the corpus. For instance, we found that we had some extra

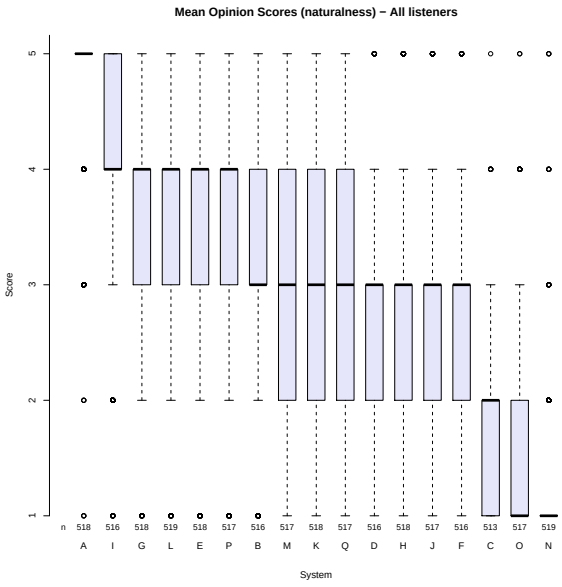

Figure 3: *Mean Opinion Scores, naturalness evaluation, all listeners.*

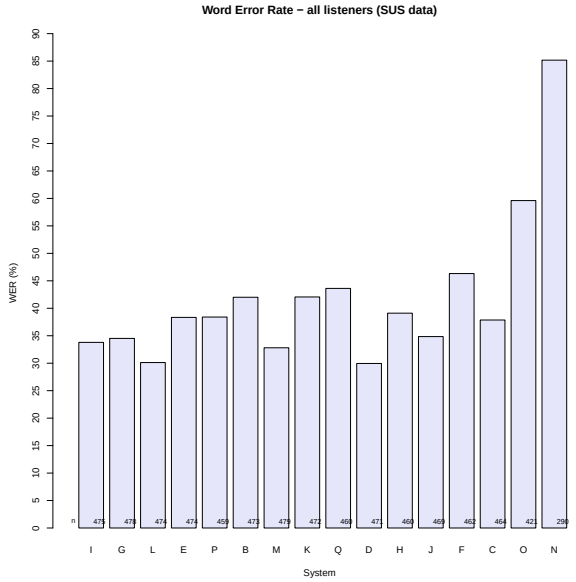

Figure 4: *Word Error Rates, intelligibility evaluation, all listeners.*

text in some books that the segmentation system was not able to align.

Finally, other parameters as the size of the beam for the search, or the size of the candidates short list, are still difficult to tune. One important point is that those two parameters need to be chosen considering a trade-off between the number of constraints added during the unit selection and the variability of the corpus.

8. Conclusion

We described the unit-selection based IRISA system for the Blizzard challenge 2017. The unit selection method is based on a classic concatenation cost to which we add a fuzzy penalty that depends on phonological features. In order to improve the system, we added specific costs to deal with prosody and transitions between narrative third-person and first person text. Despite the improvements we have made, our system obtained average results. One explanation is that by using the narrative/direct speech sub-cost, we added to many constraints during the unit selection process leading to inconsistencies in stress and prosody. Bad stress placement is also the result of the relaxation of stress constraints when it should not be the case. These two elements were the cause of a drop in nearly all criteria.

9. Acknowledgements

This study has been partially funded thanks to the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015.

10. References

[1] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proc. of Interspeech*, 2008, pp. 2–5.

[2] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7962–7966.

[3] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. of ICASSP*. IEEE, 1988, pp. 679–682.

[4] A. W. Black and P. Taylor, "CHATR: a generic speech synthesis system," in *Proc. of Coling*, vol. 2. Association for Computational Linguistics, 1994, pp. 983–986.

[5] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of ICASSP*, vol. 1. Ieee, 1996, pp. 373–376.

[6] P. Taylor, A. 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.

[7] A. P. Breen and P. Jackson, "Non-uniform unit selection and the similarity metric within BTs Laureate TTS system," in *Proc. of the ESCA Workshop on Speech Synthesis*, 1998, pp. 373–376.

[8] R. A. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.

[9] H. Patil, T. Patel, N. Shah, H. Sailor, R. Krishnan, G. Kasthuri, T. Nagarajan, L. Christina, N. Kumar, V. Raghavendra, S. Kishore, S. Prasanna, N. Adiga, S. Singh, K. Anand, P. Kumar, B. Singh, S. Binil Kumar, T. Bhadran, T. Sajini, A. Saha, T. Basu, K. Rao, N. Narendra, A. Sao, R. Kumar, P. Talukdar, P. Acharyaa, S. Chandra, S. Lata, and H. Murthy, "A syllable-based framework for unit selection synthesis in 13 indian languages," in *Proc. O-COCOSDA*, 2013, pp. pp.1–8.

[10] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," *Proc. of ICASSP*, vol. 2, pp. 837–840, 2001.

[11] D. Tihelka, J. Matoušek, and Z. Hanzlíček, "Modelling F0 Dynamics in Unit Selection Based Speech Synthesis," in *Proc. of TSD*, 2014, pp. 457–464.

[12] J. Yi, "Natural-sounding speech synthesis using variable-length units," Ph.D. dissertation, 1998.

[13] D. Cadic, C. Boidin, and C. D'Alessandro, "Vocalic sandwich, a unit designed for unit selection TTS," in *Proc. of Interspeech*, no. 1, 2009, pp. 2079–2082.

[14] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[15] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech coding and synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier Science, 1995, pp. 495–518.

[16] J. Duddington, "eSpeak text to speech," 2012.

[17] S. Young, G. Evermann, M. Gales, T. Hein, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev *et al.*, "The HTK book. for version 3.3 (april 2005)," 2013.

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

[19] D. Guennec and D. Lolive, "Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System," in *Proc. of TSD*, 2014, pp. 432–440.

[20] P. Alain, J. Chevelu, D. Guennec, G. Lecorvé, and D. Lolive, "The IRISA Text-To-Speech System for the Blizzard Challenge 2016," in *Blizzard Challenge 2016 workshop*, Cupertino, United States, Sep. 2016.

[21] C. Blouin, O. Rosec, P. Bagshaw, and C. D'Alessandro, "Concatenation cost calculation and optimisation for unit selection in TTS," in *IEEE Workshop on Speech Synthesis*, 2002, pp. 0–3.

[22] F. Alías, L. Formiga, and X. Llorá, "Efficient and reliable perceptual weight tuning for unit-selection text-to-speech synthesis based on active interactive genetic algorithms: A proof-of-concept," *Speech Communication*, vol. 53, no. 5, pp. 786–800, May 2011.

[23] D. Guennec and D. Lolive, "On the suitability of vocalic sandwiches in a corpus-based tts engine," in *Proc. of Interspeech*, 2016.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.