Improving TTS with corpus-specific pronunciation adaptation

Marie Tahon, Raheel Qader, Gwénolé Lecorvé, Damien Lolive

Improving TTS with corpus-specific pronunciation adaptation

Marie Tahon[1], *Raheel Qader*[1], *Gwénolé Lecorvé*[1], *Damien Lolive*[1]

[1]IRISA/University of Rennes 1, Lannion, France

{marie.tahon, raheel.qader, gwenole.lecorve, damien.lolive}@irisa.fr

Abstract

Text-to-speech (TTS) systems are built on speech corpora which are labeled with carefully checked and segmented phonemes. However, phoneme sequences generated by automatic grapheme-to-phoneme converters during synthesis are usually inconsistent with those from the corpus, thus leading to poor quality synthetic speech signals. To solve this problem, the present work aims at adapting automatically generated pronunciations to the corpus. The main idea is to train corpus-specific phoneme-to-phoneme conditional random fields with a large set of linguistic, phonological, articulatory and acoustic-prosodic features. Features are first selected in cross-validation condition, then combined to produce the final best feature set. Pronunciation models are evaluated in terms of phoneme error rate and through perceptual tests. Experiments carried out on a French speech corpus show an improvement in the quality of speech synthesis when pronunciation models are included in the phonetization process. Appart from improving TTS quality, the presented pronunciation adaptation method also brings interesting perspectives in terms of expressive speech synthesis.

Index Terms: speech synthesis, conditional random fields, pronunciation adaptation, feature selection.

1. Introduction

The objective of speech synthesis is to generate a speech waveform from an input text. To do so, among other processings, this text is converted into a phoneme sequence using a phonetizer. Then, the waveform is generated from this phoneme sequence by querying a dedicated database of speech segments or generative models, be it a unit selection or a statistical parametric text-to-speech (TTS) system. In both cases, the system has been built using a speech corpus in which realized phonemes have been carefully labeled and segmented. Hence, TTS systems highly depend on the consistency between phonemes as labeled in their underlying speech corpus and those generated by the phonetizer during synthesis. Especially, strong differences would lead in a low quality of the synthetized speech signals. In the case of unit selection, inconsistencies would result to a low number of candidate segments and a high number of concatenations, while, in systems like HTS, they would end up in using poorly trained or non-contextual models. To solve this problem, this paper proposes a new pronunciation adaptation method which adapts phonemes generated by the phonetizer to the speech corpus.

While voice technologies are expanding rapidly, natural language processing (NLP) systems generally rely on a very small variety of voices, thus leading to culturally centered and neutrally accented systems [1]. Therefore, one of the current challenge in NLP and more specifically in speech synthesis is the adaptation of models to a specific expressivity, a speaking style, or to speaker characteristics [2]. A possible way to introduce pronunciation variants into TTS is to manually add alternative pronunciations directly into the dictionary [3, 4]. In recent litterature, machine learning and statistical approaches have been proposed in both automatic speech recognition (ASR) and TTS systems. For example, in Karanasou et al. [5], a phoneme confusion model is trained using neural networks and conditional random fields (CRF) models. CRF and weighted finite transducers have also been used to generate probabilistic pronunciation lattices [6, 7]. Articulatory features trained with dynamic Bayesian networks have been shown to be relevant for pronunciation modeling [8].

Many features have been used to train pronunciation models. Linguistic, phonological and articulatory features are derived from textual data, such as distinction between content and function words, word predictability, syllable locations [9, 10, 11]. Syllable-based features have been investigated for pronunciation variations in French [12]. Articulatory features describe physiological properties of the speech production process. Articulatory features have been shown to be relevant for pronunciation modeling [13]. Acoustic features (mainly cepstral features) have also been used to study variations in pronunciation in ASR [14]. In TTS, some prosodic features can be extracted with a text-to-prosody model. Chen and Hasegawa-Jonhson [15] showed that prosodic features affect pronunciation particularly for spontaneous speech. Two specificities of French language (schwa and liaisons) have been shown to be very important cues in pronunciation variants [16].

The present paper improves the method proposed in [17] and adapts it to a French speech corpus. 52 linguistic, phonological, articulatory and prosodic features are first selected with a forward selection algorithm in cross-validation condition, then combined to produce the final best feature set. Pronunciation models are evaluated in terms of phoneme error rate and through perceptual tests. The obtained results confirms that corpus-specific pronunciation adaptation improves TTS quality.

In the remainder, the speech corpus, its derived features and the experimental set-up are introduced in Section 2. Feature selection protocol and results are presented in Section 3. The proposed pronunciation adaptation method is finally evaluated through phoneme error rates and perceptual tests in Section 4. Conclusion and perspectives are drawn in the last section.

2. Material and method

This section is devoted to the presentation of the speech corpus used in the experiments, the description of the feature set and the presentation of the experimental set-up.

2.1. Speech corpus

Experiments were carried out on a French speech corpus dedicated to interactive vocal system TTS. As such, this corpus covers all diphonemes present in French and comprises most

Table 1: *Groups of features used for pronunciation modeling experiments. In bold, features that have been selected. In brackets, the number of votes [nv].*

Linguistic features (18)
Word [7] ♦ **Stem** [7] ♦ Lemma [0] ♦ POS [2] ♦ Stop word [0] ♦ Word [0], stem [2], lemma [1] freq. in French (common, normal, rare) ♦ Word [1], stem [1], lemma [2] freq. in corpus ♦ Word freq. knowing previous word in French [2], in corpus [1] ♦ Word freq. knowing next word in French [2] in corpus [3] ♦ Number of word occurence in corpus [0] (numerical) ♦ Word position [3], reverse position [0] in utterance (numerical)
Phonological features (17)
Canonical syllables [7] ♦ Phoneme in syllable position [0] ♦ Phoneme in word position [0] (begin, middle, end) ♦ **Syllable in word position** [6] ♦ Phoneme position [0] and **reverse position** [4] in syllable (numerical) ♦ **Phoneme position** [5] and **reverse position** [5] in word (numercial) ♦ Syllable position [3] and reverse position [1] in word (numercial) ♦ **Word length in phoneme** [4] (numerical) ♦ Word length in syllable [2] (numerical) ♦ Syllable short [1] and long [0] structure (CVC, CCVCC) ♦ Syllable type [1] (open, closed) ♦ Phoneme in syllable part [0] (onset, nucleus, coda) ♦ **Pause per Syllable** [4] (low, normal, high)
Articulatory features (9)
Phoneme type [2] (vowel, consonant) ♦ Phoneme aperture [3], shape [1], place [1] and manner [2] (open, close, front, central, undef, etc.) ♦ Phoneme is affricate [0], rounded [3], doubled [0] or voiced [3] ? (boolean)
Prosodic features (7)
Syllable Energy [7] (low, normal, high) ♦ **Syllable [4] and phoneme [7] tone** (from 1 to 5) ♦ F_0 **phoneme contour** [7] (decreasing, flat, increasing) ♦ **Speech rate** [7] (low, normal, high) ♦ Distance to next [3] and **previous pause** [7] (from 1 to 3)

used words in the telecommunication field. It features a neutral female voice sampled at 16kHz (lossless encoding, one channel). The corpus is composed of $7,208$ utterances, containing $225,08$ phonemes and $24,160$ non speech sounds, totaling 6h40' of speech. Pronunciations and non speech sounds have been strongly controled during the recording process. Other information has been automatically added and manually corrected. The corpus and its annotations are managed using the Roots toolkit [18].

2.2. Features

The goal of the present work is to reduce the differences between phonemes generated by the phonetizer during synthesis, referred to as *canonical phonemes*, and phonemes as labeled in the speech corpus, referred to as *realized phonemes*. To do so, the proposed method is to train a CRF model which predicts corpus-specific phonemes from canonical ones. To enrich the model, and hopefully improve the prediction accuracy, other state-of-the-art features are added. Precisely, four groups of features have been investigated: linguistic, phonological, articulatory and prosodic features. The corresponding set of 52 feature presented in Table 1 is inspired from [17]. It has been enriched and adapted to French. Most features have been normalized to corpus or utterance and discretized.

Canonical phonemes are generated with Liaphon [19], one of the most widely used utterance phonetization system for French. Word frequencies in French are extracted from Google ngrams [20]. Articulatory features are standard International Phonetic Alphabet (IPA) traits. In an ideal system, prosody should also be predicted from text. However, because this task is still a research issue, prosodic features have been extracted in a oracle way, i.e., directly from the recorded utterances of the speech corpus. Such a protocol allows us to know in what extent prosody affects pronunciation models. Prosodic features are based on energy, fundamental frequency F_0 and duration. F_0 shape is based on a glissando value perceptually defined [21].

2.3. Experimental set-up

In the presented work, phonemic sequences are modeled using CRFs. They are trained using the Wapiti toolkit [22] with default BFGS algorithm. Phoneme sequences labeled by the different models are compared to the realized phoneme sequence under the usual phoneme error rate (PER).

The speech corpus has been randomly split in two: a training set (70%) and a validation set (30%). The training set has been divided in seven folds, and used to select and combine features in cross-validation conditions. Models are trained on six folds, the remaining fold being used for testing. The validation set is used to evaluate the resulting pronunciation models in final experiments in terms of PER and through perceptual tests. This protocol ensures that data used for training the models and data used for validation do not overlap.

3. Cross-validation feature selection

Feature selection is a very important task in machine learning. It helps to identify the feature subset which best predicts pronunciation, usually avoids overfitting the training data, and thus leads to models that generalize more to unseen data. Lastly it reduces the time and memory required during the training process. In our method, features are selected separately for each group of features using a forward selection process. Then groups of selected features are combined to find the optimal configuration.

3.1. Forward selection protocol

For each group of features, the forward feature selection starts with canonical phonemes only and other features are added one at a time until the optimal subset is reached. This optimal subset is found when the addition of one more feature does not improve the PER, modulo a fixed ϵ value:

$$PER(n+1) > PER(n) - \epsilon \,,$$

where n is the number of features. In practice, ϵ has been empirically set to 0.1. In order to find the global subset from the seven subsets obtained for each fold, a voting process has been

Table 2: *Average PERs on the training set obtained on 7 folds. In brackets, percentage point w.r.t. the baseline.*

Baseline (no adaptation)		11.5 [0.0]
Canonical phoneme only (C)		6.9 [-4.6]
Linguistic (C+L)	all (18)	4.4 [-7.1]
	selected (2)	4.4 [-7.1]
Phonological (C+Ph)	all (17)	4.5 [-7.0]
	selected (7)	4.6 [-6.9]
Articulatory (C+A)	all (9)	7.1 [-4.4]
	selected (0)	-
Prosodic (C+Pr)	all (7)	4.8 [-6.7]
	selected (6)	4.8 [-6.7]
C + L + Ph	selected (9)	4.0 [-7.5]
C + L + Pr	selected (8)	3.5 [-8.0]
C + Ph + Pr	selected (13)	3.6 [-7.9]
C + L + Ph + Pr	selected (15)	3.2 [-8.3]

set up. For each fold, a selected feature receives a vote $v = 1$, then the maximum of votes for the global selection process is $nv = 7$. Features which receive a number of votes $nv \geq 4$ (i.e. $> 50\%$), are added in the global subset.

3.2. Selected features

Selected features are reported in bold in Table 1 along with their number of votes nv. First, it appears that two linguistic features were selected for all folds: the word itself and its stem. Since these features are highly correlated, one would have expected only one feature to be selected. However, as stated in [23], "noise reduction and consequently better class separation may be obtained by adding variables that are presumably redundant". Word frequencies and left/right linguistic context features, received only very few votes. Surprisingly, it appears that no articulatory features has reached the minimal number of votes. Since previous studies have shown the interest of such features for pronunciation variation modeling [8], they were expected to have better votes. Then, seven phonological features were included in the optimal set. Most of the selected features concern phoneme positions in the utterance. None of the characteristics of syllables (such as syllable part, structure or type) have been selected. Finally, six out of seven prosodic features have been selected. Five of them reached the maximum number of votes. This result is in agreement with state-of-the-art and suggests that a prosodic model should improve speech synthesis.

Average PER obtained on the seven folds are reported in Table 2. The baseline is the PER obtained without any adaptation, between phoneme sequences generated by the phonetizer and realized phoneme sequences (ground truth). An improvement of 4.6 percentage point (pp) is obtained while using a pronunciation model trained with canonical phonemes only, thus showing how pronunciation adaptation can reduce the inconstancy between the phonetizer output and the speech corpus. Separately adding group of features further improves the PER, except with the articulatory group. Interestingly, the reduction of the number of features in each group does not affect these average PERs. The most spectacular reduction lies in the linguistic group: with only two apparently redundant features, a drop of 7.1 pp is obtained from the baseline.

Table 3: *PERs obtained on the test set. In brackets, percentage point w.r.t. the baseline.*

Baseline (no adaptation)		11.2 [0.0]
Canonical phoneme only (C)		6.6 [-4.6]
C + L + Ph	selected (9)	3.9 [-7.3]
C + L + Ph + Pr	selected (15)	3.3 [-7.9]

3.3. Feature groups combination

Once feature selection is performed for each goup, the combination of these groups has been investigated to find the optimal configuration. Table 2 summarizes PERs of all possible combinations of selected feature groups. Overall results show an improvement in PER when combining groups. The combination of prosodic and linguistic groups leads to a significant drop in PER of 8.0 pp with a minimum number of features. The combination of the three feature groups brings the best PER, with an improvement of 8.3 pp from the baseline. In the end, only almost a third of the initial feature set remains.

4. Evaluation

This section focuses on the validation of the conclusions obtained in the previous Section. Pronunciation models are now trained on the whole training set and tested on the validation set. Pronunciation models are tested with different subsets of feature: canonical phonemes only, best selected linguistic and phonological features and best selected linguistic, phonological and prosodic features. The study of how the prosodic features affect the results is of particular importance in the context of TTS.

4.1. Phoneme sequence validation

The obtained results are almost the same as in Section 3, the best configuration being the combination of selected features including prosodic features (- 7.9 pp). However the improvement brought by prosody is not as important here (0.4 pp) as it was in the training step (0.8 pp). Then, linguistic and phonological features are probably more robust to unseen data than prosodic features. Based on the PER results, we are expecting that the addition of linguistic, phonological and prosodic features improves the synthesized speech quality.

Most confusions between canonical and realized phonemes concern allophones: o ≒ ɔ, e ≒ ɛ and ɛ̃ ≒ œ̃. Such confusions cannot be considered as errors in French. They depend on the speaking style. Similarly frequent reported insertions concern the ə which is known to be optionaly elided. Other substitutions concern labeling strategies and alphabet choices, for example ɲ ≒ nj, ə ≒ ø. Deletions mainly concern liaisons between words, such as t, z which are not generated by the phonetizer whereas systematically pronounced in the speech corpus. Pronunciation models contribute to minimize all these confusions.

4.2. Synthesized speech evaluation

In order to assess the quality of synthesized speech samples generated with adaped pronunciation, a perceptual test was conducted with 14 French native speakers. The evaluation is based on AB tests with 40 utterances in which listeners have to answer the following question: *"Between A and B, which sam-*

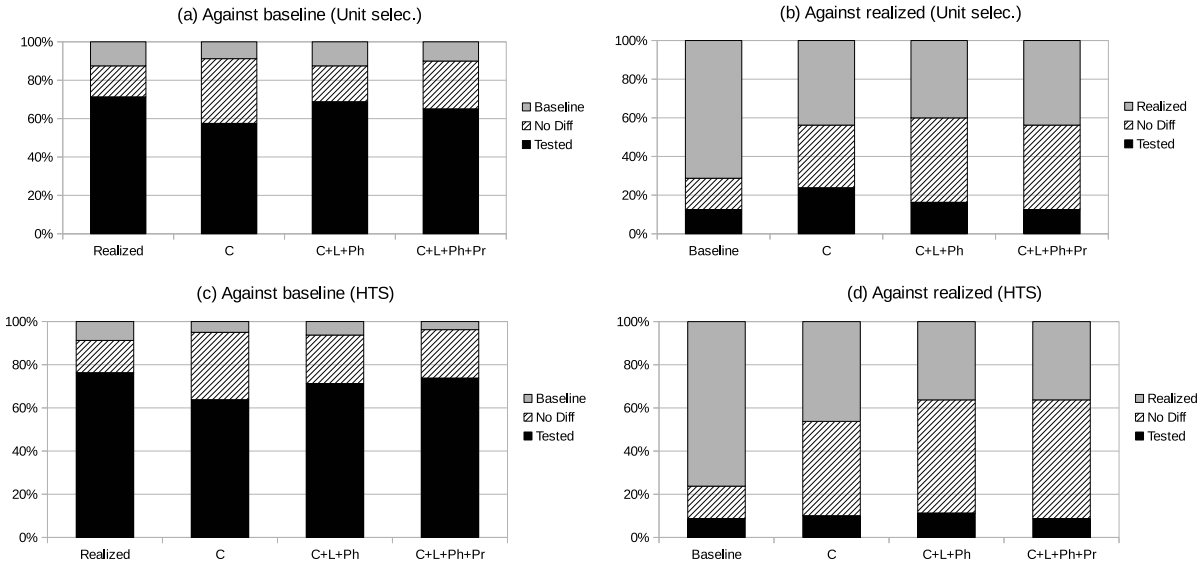

Figure 1: *AB test results with unit selection (a,b) and HTS (c,d): number of times a system is chosen. up: {realized, C, C+L+Ph, C+L+Ph+Pr} against baseline (a,c), down: {baseline, C, C+L+Ph, C+L+Ph+Pr} against realized (b,d)*

ple reaches the best quality ?". Possible answers are: *A*, *B*, or *no difference*. Utterances were randomly selected by subsampling the validation set according to the PER distribution between canonical and realized pronunciations. Speech samples were synthesized using the corpus-based TTS system described in [24] and also with HTS v2.2 with standard features [25]. Some examples extracted from the listening test are available on the team website[1]. Five pronunciations are evaluated: canonical phonemes without adaptation (baseline), adapted phonemes based on canonical phonemes (C), selected linguistic and phonological features (C+L+Ph), selected linguistic, phonological and prosodic features (C+L+Ph+Pr) and realized phonemes as they are annotated in the speech corpus.

Fig. 1 shows the comparison of speech samples using adaptation against the baseline (left) and the realized (right) pronunciations with the two synthesis systems. Tested systems are expected to be mostly preferred against the baseline, that is the larger the black bar, the better. At the opposite, the tested system is considered as correct when its signals are preferred or judged as similar against the realized signals, that is the smaller the grey bar, the better. With both synthesis systems, adapted pronunciations resulting from the presented approach outweighs the baseline pronunciations in terms of quality. The addition of linguistic and phonological increases the number of preferred adapted pronunciations. However, prosodic features do not seem to improve TTS quality, what is of interest since these features are not easy to obtain from text.

The adapted pronunciations can be considered as correct in comparison to realized pronunciations because the synthesized adapted pronunciation are mainly judged as similar or even better to the realized pronunciation (in more than 50% of the samples). Interestingly, the C+L+Ph configuration is even more preferred than the configuration with prosodic features. This confirms that linguistic and phonological features are more robust than prosodic features. Based on this perceptual evaluation, it seems that pronunciation adaptation using linguistic and phonological features is our best model.

[1] https://www-expression.irisa.fr/demos/

5. Conclusions and perspectives

In this paper, we have presented a new pronunciation adaptation method which adapts phonemes generated by the phonetizer to the speech corpus. A CRF pronunciation model trained with linguistic, phonological and prosodic features have been proposed. Feature were selected using a forward feature selection algorithm in cross-validation configuration, thus reducing the initial feature set from 52 to 15 features. Articulatory features were never selected.

The proposed corpus-specific pronunciation adaptation method brings an improvement of 7.9 pp in terms of phoneme error rate. Perceptual tests also show an improvement in the quality of speech synthesis when pronunciation models are included in the phonetization process. Hence, we have shown that pronunciation adaptation helps to reduce inconsistencies between phonemes as labeled by their underlying speech corpus and those generated by the phonetizer during synthesis. Moreover, one of the advantages of this statistical approach is to be easily reproductible.

Further experiments are needed to improve pronunciation adaptation models. For example, considering previous and next canonical phonemes could lead to better results on liaisons. Apart from improving TTS quality, the presented pronunciation adaptation method also brings interesting perspectives in terms of expressive speech synthesis. The use of n-best output phonemes predicted with CRF probabilities for extracting pronunciation lattices should improve many TTS applications where a specific expressivity or speaking style is needed.

6. Acknowledgements

This study has been realized under the ANR (French National Research Agency) project SynPaFlex ANR-15-CE23-0015.

7. References

[1] C. Olinsky and F. Cummins, "Iterative English adaptation in a speech synthesis system," in *IEEE Workshop on Speech Synthesis*, Sept. 2002, pp. 79–82.

[2] D. Govind and S. M. Prasanna, "Expressive speech synthesis: a review," *International Journal of Speech Technology*, vol. 16, pp. 237–260, 2013.

[3] T. Fukada, T. Yoshimura, and Y. Sagisaka, "Automatic generation of multiple pronunciation based on neural networks." *Speech Communication*, vol. 27, pp. 63–73, 1999.

[4] G. Tajchman, E. Foster, and D. Jurafsky, "Building multiple pronunciation models for novel words using exploratory computational phonology," in *European Conference on Speech Communication and Technology (Eurospeech)*, 1995.

[5] P. Karanasou, F. Yvon, T. Lavergne, and L. Lamel, "Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR," in *Annual Conference of the International Speech Communication Association (Interspeech)*, Lyon, France, August 2013, pp. 1966–1970.

[6] G. Lecorvé and D. Lolive., "Adaptive statistical utterance phonetization for French." in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 4864–4868.

[7] T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *Speech Communication*, vol. 46, no. 2, pp. 189–203, 2005.

[8] K. Livescu, P. Jyothi, and E. Fosler-Lussier, "Articulatory feature-based pronunciation modeling," *Computer Speech and Language*, vol. 36, pp. 212–232, 2016.

[9] B. Vazirnezhad, F. Almasganj, and S. Ahadi, "Hybrid statistical pronunciation models designed to be trained by a medium-size corpus," *Computer Speech and Language*, vol. 23, no. 1, pp. 1–24, 2009.

[10] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea, "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation." *J. Acoust. Soc. Am.*, vol. 113, pp. 1001–1024, 2003.

[11] A. Bell, J. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, no. 1, pp. 92–111, 2009.

[12] M. Adda-Decker, P. B. de Mareüil, G. Adda, and L. Lamel, "Investigating syllabic structures and their variation in spontaneous French," *Speech Communication*, vol. 46, pp. 119–139, 2005.

[13] R. A. Bates, M. Osendorf, and R. A. Wright, "Symbolic phonetic features for modeling of pronunciation variation," *Speech Communication*, vol. 49, pp. 83–97, 2007.

[14] C. L. Bennett and A. W. Black, "Prediction of pronunciation variation for speech synthesis: a data-driven approach," in *ICASSP*, 2005, pp. 297–300.

[15] K. Chen and M. Hasegawa-Johnson, "Modeling pronunciation variation using artificial neural networks for English spontaneous speech," in *8th International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, Korea, October 2004, pp. 4–8.

[16] P. B. de Mareüil and M. Adda-Decker, "Studying pronunciation variants in French by using alignment techniques." in *International Conference on Spoken Language Processing*, Denver, Colorado, USA, September 2002.

[17] R. Qader, G. Lecorvé, D. Lolive, and P. Sébillot, "Probabilistic speaker pronunciation adaptation for spontaneous speech synthesis using linguistic features," in *International Conference on Statistical Language and Speech Processing (SLSP)*, 2015.

[18] J. Chevelu, G. Lecorvé, and D. Lolive, "ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections," in *9th International Language Resources and Evaluation Conference (LREC), European Language Resources Association (ELRA)*, Reykjavik, Iceland, May 2014.

[19] F. Béchet, "LIA-PHON: un système complet de phonétisation de texte," *Traitement Automatique des Langues (TAL)*, vol. 42, no. 1, pp. 47–67, 2001.

[20] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Republic of Korea, July 2012, pp. 169–174.

[21] C. dAlessandro, S. Rosset, and J.-P. Rossi, "The pitch of short-duration fundamental frequency glissandos," *J. Acoust. Soc. Am.*, vol. 104, pp. 2339–2348, 1998.

[22] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden, 2010, pp. 504–513.

[23] I. Guyon and A. Elissef, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[24] 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.

[25] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." in *Speech Synthesis Workshop (SSW)*, 2007, pp. 294–299.