Statistical Pronunciation Adaptation for Spontaneous Speech Synthesis

Raheel Qader, Gwénolé Lecorvé, Damien Lolive, Marie Tahon, Pascale Sébillot

Statistical Pronunciation Adaptation for Spontaneous Speech Synthesis

Raheel Qader[1], Gwénolé Lecorvé[1], Damien Lolive[1], Marie Tahon[1], and Pascale Sébillot[2]

[1] IRISA/University of Rennes 1 (ENSSAT), Lannion, France
[2] IRISA/INSA Rennes, Rennes, France
first_name.last_name@irisa.fr

Abstract. To bring more expressiveness into text-to-speech systems, this paper presents a new pronunciation variant generation method which works by adapting standard, i.e., dictionary-based, pronunciations to a spontaneous style. Its strength and originality lie in exploiting a wide range of linguistic, articulatory and prosodic features, and in using a probabilistic machine learning framework, namely conditional random fields and phoneme-based n-gram models. Extensive experiments on the Buckeye corpus of English conversational speech demonstrate the effectiveness of the approach through objective and perceptual evaluations.

Keywords: speech synthesis, spontaneous speech, pronunciation modeling, statistical adaptation, conditional random field

1 Introduction

Modeling pronunciation variation in spontaneous speech is critical to achieve expressive Text-To-Speech (TTS) synthesis since pronunciation variants reflect the emotional state of a speaker, his/her intention, or a specific accent. However, phonetizers used by most current TTS systems fail to capture these variants as they only rely on standard pronunciations, i.e., extracted or learned from a general dictionary. Thus, the resulting synthetic speech conveys a neutral and formal style. A solution to this problem is to adapt standard pronunciations in order to reflect spontaneousness. In a machine learning perspective, this task corresponds to predicting a sequence of spontaneous phonemes from an input sequence of canonical phonemes, i.e., deciding whether input phonemes should be deleted, substituted, simply kept as is, or if new phonemes should be inserted.

Most of the early work in the area of pronunciation adaptation relied on using predefined or automatically extracted phonological rules to derive alternative pronunciations [1–3]. In the recent literature, various machine learning and statistical approaches have been proposed. Notably, decision trees [4, 5], random forests [6], neural networks [7, 8], hidden Markov models [9], and Conditional Random Fields (CRFs) [8, 10, 11] have been investigated. Regarding features, two categories are considered important to model pronunciation variation: linguistic-phonological features and prosodic ones. Linguistic-phonological features can be derived from textual data (POS, word predictability, lexical stress, etc.) [5, 12], while prosodic features (F0, energy, duration, etc.) can be directly extracted from speech signals or predicted from text using

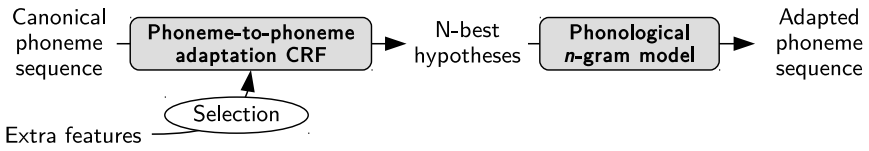

Fig. 1. Overview of the proposed pronunciation adaptation method.

a prosodic model [12, 13]. Besides those two feature types, the benefits of using articulatory features have also been experimented [14, 15]. Most of the mentioned studies have been applied in the context of Automatic Speech Recognition (ASR) and concentrated on utilizing either linguistic, articulatory or prosodic features.

In contrast, following our preliminary adaptation method proposed in [10], the method here combines a wide range of features and focuses on TTS rather than ASR. More precisely, the contributions are the following:

1. The importance and complementarity of linguistic, articulatory and prosodic information are studied w.r.t. the spontaneous style, highlighting that linguistic features are sufficient to perform good adaptations.
2. The usage of a phonological n-gram model is proposed to guarantee the *a posteriori* plausibility of the adapted pronunciations.
3. Perceptual tests demonstrate that adapted pronunciations are judged spontaneous while remaining reasonably intelligible.

In the remainder, the overal method and corpus are presented in Section 2. A study on feature selection and combination is provided in Section 3. The usage of a phonological model is exposed in Section 4. Finally, perceptual tests are discussed in Section 5.

2 Method Overview

Given a textual utterance, our fundamental idea for pronunciation adaptation is to predict the sequence of spontaneously realized phonemes from an input sequence of canonical phonemes. As shown in Figure 1, we propose to perform this task in 2 steps. First, adapted pronunciation hypotheses are generated by a phoneme-to-phoneme CRF trained on canonical phonemes and a combination of linguistic, articulatory, and prosodic features. These features are selected offline, i.e., while setting up the method, in an automatic manner to optimize the CRF accuracy. Second, hypotheses are reranked using a phonological n-gram model of spontaneous phoneme sequences.

The method is experimented on 20 hours of spontaneous American English speech from the Buckeye corpus [16]. This represents 20 interviews with speakers from central Ohio, USA, of various ages and both genders. Interviews are annotated with their orthographic transcript and two phonemic transcripts: the standard pronunciation of the words (*canonical phonemes*), and the one effectively uttered by the speaker (*realized phonemes*). The average numbers of phonemes and words per speaker are 22,789 and 7,354, respectively. The Phoneme Error Rate (PER) between the canonical and realized phonemes is $28.3\,\%$. This very high rate shows how different standard and spontaneous pronunciations are, and how difficult adapting pronunciation to a spontaneous style is.

Linguistic features (22)
canonical phoneme • **word** • **is a stop word** • **syllable lexical stress** • **syllable part** • **word frequency in English** • **reverse phoneme position in syllable** • **phoneme position in syllable** • **syllable location** • stem frequency in the interview • word frequency in the interview • syllable type • POS • number of syllables of the word • stem frequency in English • grapheme • word length • reverse utterance position • utterance position • word position • reverse word position • word occurrence count in interview
Articulatory features (9)
vowel/consonant • **manner** • **place** • **shape** • **aperture** • **voiced** • **rounded** • affricate • doubled
Prosodic features (10)
syllable energy • **syllable F0 shape** • **syllable tone** • **speech rate** • **pause per syllable** • phone tone • distance to previous silence • distance to next silence • distance to previous hesitation (*um*/*uh*) • distance to next hesitation (*um*/*uh*)

Table 1. List of all features. Selected features are in bold.

Phone segmentation is also available and about 40 linguistic-phonological (shortened to *linguistic* in the remainder), articulatory, and prosodic features have been automatically added using speech and natural language processing tools (see Table 2). Prosodic features have been directly estimated in an oracle way by processing signals of each speaker, normalizing and strongly approximating the derived information. This simulates a perfect prosody modeling, leading to adaptation results which are not biased by prosody prediction errors, while remaining realistic. Finally, the corpus has been randomly divided into a training set (60% of the utterances), a development set (20%), and a test set (20%), with an equal representation of each speaker in each set.

Phoneme sequences generated by our method are evaluated by PERs w.r.t. the ground truth, i.e., the sequence realized by the speaker. Thus, the lower the PER the better, the baseline being the PER of the canonical pronunciation, that is before adaptation. Listening tests have also been conducted to perceptually validate the method. All models have been learned on the training set, optimized on the development set and evaluated on the test set. Canonical phonemes have been automatically aligned with realized phonemes using `m2m-aligner` [17] to train the phoneme-to-phoneme CRF.

3 Phoneme-to-Phoneme Adaptation

Phoneme-to-phoneme adaptation is performed by CRFs trained on the canonical phonemes and relevant selected features. This section briefly describes how linguistic, articulatory, and prosodic features have been selected, before presenting how the selected features have been combined to produce the final adaptation CRF.

Automatic selection is applied on each feature group separately in order to eliminate irrelevant and redundant features. The selection process relies on a greedy approach where votes are assigned to the most influential features, that is features leading to the lowest PER when training adaptation CRFs. These CRFs are trained without contextual information to avoid large training time overheads, i.e., information about the neighbors of each canonical phoneme is disregarded. Features resulting from this selection

Baseline (not adapted)			28.3
Adapted using	Canonical phonemes only (C)		30.7 $(+2.4)$
	+ Linguistic features (C + L)	all (22)	26.6 (-1.7)
		selected (8)	25.1 (-3.2)
	+ Articulatory features (C + A)	all (9)	30.9 $(+2.6)$
		selected (7)	30.8 $(+2.5)$
	+ Prosodic features (C + P)	all (10)	27.1 (-1.2)
		selected (6)	26.7 (-1.6)

Table 2. PERs (%) on the development set with selected features vs. all features. CRFs are trained without contextual information. In brackets, variations from the baseline (in percentage points).

Baseline (not adapted)		28.3
Adapted using	C	24.2 (-4.1)
Adapted using	C + L	24.0 (-4.3)
	C + A	24.4 (-3.9)
	C + P	21.5 (-6.8)
Adapted using	C + L + A	24.0 (-4.3)
	C + L + P	**21.1** $(\mathbf{-7.2})$
	C + A + P	21.4 (-6.9)
	C + L + A + P	21.2 (-7.1)

Table 3. PERs (%) on the test set for all possible combinations. CRFs use contextual information.

are highlighted in bold in Table 1. Linguistic and prosodic information derived from syllables is particularly valuable, as well as information about word frequencies. Regarding articulatory features, the selection has less effects (only 2 discarded features), meaning that no clear dominance can be established among them. Table 2 reports the influence of feature selection on PER for CRFs trained on the development set and on each group of features independently. Results show that the selection is efficient for linguistic and prosodic features, whereas again almost useless for articulatory ones.

To optimize the method and search for potential complementarities, all possible combinations of selected features are tested. Moreover, adaptation CRFs are trained with contextual information, precisely 2 neighbors on the left and on the right, as this configuration has shown to lead to the best PER in preliminary tests. Table 3 reports PERs on the test set for these combination experiments. First, it appears that CRFs already perform rather well when solely relying on canonical phonemes (configuration C), thanks to contextual information. Then, when separately including the selected features, results show that linguistic features provide a small improvement (C + L), articulatory features bring worse results (C + A), and prosodic features (C + P) lead to a clear improvement with a reduction of 2.7 percentage points (pp) compared to the use of the sole canonical phonemes (C). Although prosodic features are extracted in an oracle way and thus lead to optimistic results, the latter result tends to show a strong relationship between prosody and pronunciation. When feature types are combined together, articulatory features bring worse results in all cases, definitely showing that they should not be considered in our method. On the contrary, results again demonstrate that linguistic and prosodic features are useful for pronunciation adaptation, bringing the best PER down to 21.1%. This conclusion has been validated by paired t-test and paired Wilcoxon test with confidence level $\alpha = 0.05$, whether these feature groups are considered individually or together.

		Before reranking	*After reranking*
Baseline (not adapted)		28.3	
Adapted using	C	24.2 (-4.1)	23.7 (-4.6)
	C + L	24.0 (-4.3)	23.7 (-4.6)
	C + L + P	21.1 (-7.2)	**20.6 (-7.7)**

Table 4. PERs (%) before and after reranking when adapting using canonical phonemes (C), linguistic (L), and prosodic features (P).

4 Phonological Rescoring and Reranking

A qualitative analysis of the adapted pronunciations shows that some phoneme sequences returned by the finally adopted CRF are very unlikely. For instance, the sequence /dt/ is rare in spontaneous English, and rather simply reduced to /d/ or /t/. To fix these imperfections, we propose as a second step to generate several pronunciation hypotheses using the adaptation CRF, and to rerank them according to scores given by a probabilistic phonological model[3]. Precisely, each hypothesis $\mathbf{h} = (p_1, \ldots, p_m)$ of m phonemes p_i is assigned a score $s(\mathbf{h})$ mixing the CRF and phonological model (PM) probabilities. This mixture is computed by a log-linear interpolation—which has been successfully used for N-best list reranking in various domains [18, 19]—, and is formulated as follows:

$$s(\mathbf{h}) = \Pr_{\text{CRF}}(\mathbf{h}) \times \Pr_{\text{PM}}(\mathbf{h})^{\alpha} \times \beta^{m} \,, \tag{1}$$

where α and β are two parameters to be optimized. The parameter β is used to prevent the phonological model from favoring short hypotheses. Finally, the hypothesis with the highest score is selected as the adapted pronunciation.

In our experiments, the phonological model is a phoneme-based n-gram model estimated on the training set using a Witten-Bell smoothing. The order n of the model as well as α and β have been optimized such that they minimize PER on the development set, and consequently set to 5, 0.48 and 0.024, respectively. Training, optimization and reranking have all been conducted using SRILM [20]. Reranking is performed on the 10 best hypotheses predicted by the adaptation CRF, as tuned on the development set.

As shown in Table 4, our reranking technique always reduces PERs. The largest reduction is 0.5 pp, achieved for both canonical phonemes (C) and linguistic+prosodic configurations (C + L + P). Alongside, phonological reranking surprisingly seems to obviate linguistic features (C against C + L). However, results of the perceptual tests (Section 5) show that this is not true. Overall, and given the difficulty of the task, results show that our whole approach is effective as it reduces PER to a large extent with significant improvements up to 7.7 pp w.r.t. the baseline.

5 Perceptual Tests

AB tests on 40 synthesized speech samples have been conducted with 10 native English speakers. Listeners were asked to answer two questions: "*Between A and B, which*

[3] CRFs allow dependencies between predicted phonemes but it appeared in preliminary work that using a separate phonological model is better to avoid overfitting the training data.

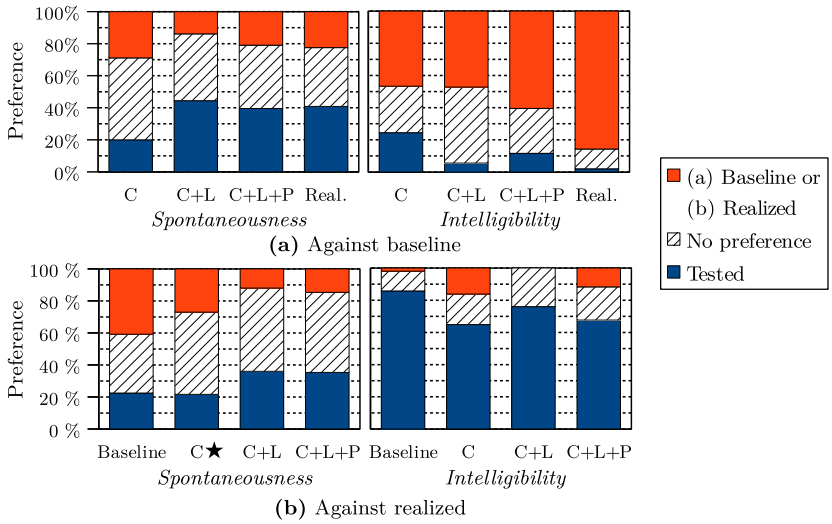

Fig. 2. Preference on spontaneousness and intelligibility between baseline, realized and adapted pronunciations. Adaptations were performed using canonical phonemes (C), linguistic features (L), and prosodic features (P). ★ stands for "*not statistically significant*[6]".

sample is pronounced in the most spontaneous *way?", and "Which once is pronounced in the most* intelligible *way?*". For both questions, listeners were also allowed to indicate that they do not have any preference. Orthographic transcripts were given along with the samples to help listeners to focus on pronunciations. Tests were set up to compare canonical and realized pronunciations to those generated by our adaptation method using either configurations C, C + L or C + L + P, all including phonological reranking.

Utterances have been selected among the 2,000 available utterances in the test set such that their PER between the canonical and realized pronunciations is high. This strategy has been designed to ensure that selected utterances reflect the difficulty of the task. Utterances were synthesized using the parametric HTS v2.2 speech synthesis system trained with standard features [21] and on the Blizzard Challenge 2012 data [22], i.e., audiobooks with mixed speech styles and uttered by a same US male speaker. Hence, no bias toward standard or spontaneous speech can be observed. Unit selection has voluntarily been discarded since this type of system is usually sensitive to pronunciation variants, producing disturbing artefacts.

Figure 2 shows the comparison of speech samples generated using (*a*) the standard pronunciations (baseline) against adapted or realized ones and (*b*) realized pronunciations against the baseline and adapted ones, in terms of spontaneousness and intelligibility. Preference percentages are given as bar segments on the y-axis. Statistical significances of these ratios have been computed for all the tests[4].

First, Figure 2.a shows that realized pronunciations are logically judged as more spontaneous than the baseline, while being much less intelligible. Regarding adapted

[4] Binomial test with $\alpha = 0.1$ and votes for "No preference" equally spread over A and B, following the methodology proposed in [23].

pronunciations, the configuration C performs poorly. Conversely, the two other adapted configurations are judged as much more spontaneous than the baseline, but again leading to intelligibility degradations. Finally, adaptation performs equally or even slightly better when using linguistic features alone, i.e., without prosodic ones. This is interesting since predicting prosodic features is difficult in TTS.

As for Figure 2.b, it surprisingly appears that C + L and C + L + P configurations are preferred over the realized pronunciations w.r.t. spontaneousness. This is probably correlated with the large intelligibility gap reported. Similarly, it can again be noticed that the use of the sole linguistic features performs slightly better than when also accounting for prosodic features, especially regarding intelligibility. On the one hand, these results demonstrate the effectiveness of our method in generating spontaneous pronunciations. On the other, they are in contradiction with PERs of Table 4. A deeper analysis shows that pronunciations produced using prosodic features, as well as the realized ones, seem to be more spontaneous but they are too complex for current TTS systems, especially because of strong coarticulations like /dn̩/ (like in "didn't") or /fm/ ("familiarity"). This penalizes intelligibility and, as a side effect, spontaneousness. Hence, a reasonable conclusion is that the proposed method is enough effective yet to reflect a spontaneous style while results could still be improved, on the condition that the speech data used to build the TTS voice are consistent with the desired degree of spontaneousness.

6 Conclusions and Future Work

In this paper, we have proposed a TTS-dedicated spontaneous pronunciation adaptation method which combines a phoneme-to-phoneme CRF and a phonological model. Objective and perceptual tests have shown that produced pronunciations effectively better reflect spontaneous speech than non adapted ones. The study of linguistic, articulatory and prosodic features shows that linguistic features are good predictors for spontaneous pronunciations, while articulatory ones are useless and prosodic ones tend to produce less intelligible speech. More generally, it seems that there is a tradeoff between the degree of spontaneousness and intelligibility. In the future, it would be interesting to more deeply investigate the relationship between prosody and pronunciation, and their impact on the perceived spontaneousness. Following this direction, finding out mechanisms to enable a fine control of intelligibility against spontaneousness is another interesting perspective, especially by taking into account the intrinsic phonemic and prosodic variability of the TTS system's voice. Finally, we have planed to apply the proposed method to emotional speech to generate synthetic speech samples.

Acknowledgments

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

References

1. Tajchman, G., Foster, E., Jurafsky, D.: Building multiple pronunciation models for novel words using exploratory computational phonology. In: Proc. of Eurospeech. (1995)

2. Giachin, E., Rosenberg, A., Lee, C.H.: Word juncture modeling using phonological rules for HMM-based continuous speech recognition. In: Proc. of ICASSP. (1990)
3. Oshika, B.T., Zue, V.W., Weeks, R.V., Neu, H., Aurbach, J.: The role of phonological rules in speech understanding research. IEEE Transactions on Acoustics, Speech and Signal Processing **23** (1975)
4. Goronzy, S., Rapp, S., Kompe, R.: Generating non-native pronunciation variants for lexicon adaptation. Speech Communication **42**(1) (2004)
5. Vazirnezhad, B., Almasganj, F., Ahadi, S.M.: Hybrid statistical pronunciation models designed to be trained by a medium-size corpus. Computer Speech & Language **23** (2009)
6. Dilts, P.C.: Modelling phonetic reduction in a corpus of spoken English using random forests and mixed-effects regression. PhD thesis, University of Alberta (2013)
7. Chen, K., Hasegawa-Johnson, M.: Modeling pronunciation variation using artificial neural networks for English spontaneous speech. In: Proc. of Interspeech. (2004)
8. Karanasou, P., Yvon, F., Lavergne, T., Lamel, L.: Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR. In: Proc. of Interspeech. (2013)
9. Prahallad, K., Black, A.W., Mosur, R.: Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis. In: Proc. of ICASSP. (2006)
10. Qader, R., Lecorvé, G., Lolive, D., Sébillot, P.: Probabilistic speaker pronunciation adaptation for spontaneous speech synthesis using linguistic features. In: Proc. of SLSP. (2015)
11. Tahon, M., Qader, R., Lecorvé, G., Lolive, D.: Improving TTS with corpus-specific pronunciation adaptation. In: Proc. of Interspeech. (2016)
12. Bell, A., Brenier, J.M., Gregory, M., Girand, C., Jurafsky, D.: Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language **60** (2009)
13. Bates, R., Ostendorf, M.: Modeling pronunciation variation in conversational speech using prosody. In: Proc. of ISCA Tutorial and Research Workshop (ITRW) on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology. (2002)
14. Livescu, K., Jyothi, P., Fosler-Lussier, E.: Articulatory feature-based pronunciation modeling. Computer Speech & Language **36** (2016)
15. Rasipuram, R., Doss, M.M.: Articulatory feature based continuous speech recognition using probabilistic lexical modeling. Computer Speech & Language **36** (2016)
16. Pitt, M.A., Johnson, K., Hume, E., Kiesling, S., Raymond, W.: The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. Speech Communication **45** (2005)
17. Jiampojamarn, S., Kondrak, G., Sherif, T.: Applying many-to-many alignments and hidden markov models to letter-to-phoneme conversion. In: Proc. of NAACL-HLT. (2007)
18. Rosti, A.V.I., Matsoukas, S.: Combining outputs from multiple machine translation systems. In: Proc. of NAACL-HLT. (2007)
19. Huet, S., Gravier, G., Sébillot, P.: Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition. Computer Speech & Language **24**(4) (2010)
20. Stolcke, A., Zheng, J., Wang, W., Abrash, V.: SRILM at sixteen: Update and outlook. In: Proc. of IEEE ASRU Workshop. (2011)
21. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The HMM-based speech synthesis system (HTS) version 2.0. In: Proc. of SSW. (2007)
22. King, S., Karaiskos, V.: The Blizzard Challenge 2012. In: Proc. of Blizzard Challenge 2012 Workshop. (2012)
23. Karhila, R., Remes, U., Kurimo, M.: Noise in HMM-based speech synthesis adaptation: analysis, evaluation methods and experiments. IEEE Journal of Selected Topics in Signal Processing **8**(2) (2014)