Optimal feature set and minimal training size for pronunciation adaptation in TTS

Marie Tahon*, Raheel Qader, Gwénolé Lecorvé, and Damien Lolive

IRISA/University of Rennes 1
6 rue de Kérampont, 22300 Lannion, France
{marie.tahon,raheel.qader,gwenole.lecorve,damien.lolive}@irisa.fr
https://www-expression.irisa.fr/

Abstract. Text-to-Speech (TTS) systems rely on a grapheme-to-phoneme converter which is built to produce canonical, or statically stylized, pronunciations. Hence, the TTS quality drops when phoneme sequences generated by this converter are inconsistent with those labeled in the speech corpus on which the TTS system is built, or when a given expressivity is desired. To solve this problem, the present work aims at automatically adapting generated pronunciations to a given style by training a phoneme-to-phoneme conditional random field (CRF). Precisely, our work investigates (i) the choice of optimal features among acoustic, articulatory, phonological and linguistic ones, and (ii) the selection of a minimal data size to train the CRF. As a case study, adaptation to a TTS-dedicated speech corpus is performed. Cross-validation experiments show that small training corpora can be used without much degrading performance. Apart from improving TTS quality, these results bring interesting perspectives for more complex adaptation scenarios towards expressive speech synthesis.

Keywords: Speech synthesis, pronunciation adaptation, feature selection, training data size.

1 Introduction

Text-to-speech (TTS) systems mainly rely on two steps. First, the input text is converted to a canonical phoneme sequence using an automatic phonetizer. Then, the waveform is generated from this phoneme sequence by querying a dedicated database of speech segments or using generative models trained on this database. In such a framework, and as used in current TTS systems, phonemes generated by the phonetizer need to be consistent with those labeled in the speech corpus in order to produce high quality synthetic speech samples. Given an existing TTS application, this strong requirement makes it very difficult to change the phonetizer to another, move to a new speech database, or to consider expressive pronunciation variants, unless redesigning all the components from scratch. As a consequence, TTS applications are nowadays still grounded on

* Corresponding author

a very limited variety of voices, yielding to culturally centered and neutrally accented systems [1]. One of the current challenges in TTS is thus to adapt standard pronunciations to new conditions, especially expressivity, speaking style or speaker characteristics [2].

To overcome this problem, our paper focuses on a new pronunciation adaptation method which adapts canonical phonemes generated by the phonetizer to a specific pronunciation style. Precisely, this method seeks here to adapt canonical phonemes to pronunciations uttered in the speech corpus on which the TTS system is built. Beyond the importance of this particular problem, our work is more generally regarded as a case study towards more complex adaptation scenarios, e.g., emotion or accent-specific adaptations. Using machine learning, the studied pronunciation adaptation method consists in training adaptation models on a target *pronunciation corpus*. In the perspective to deploy this method to various use cases, investigations are conducted in this paper on (i) the choice of optimal features and (ii) the minimal size of the pronunciation corpus to train reasonable adaptation models. Looking for an optimal feature set to model pronunciations is required to improve the adaptation accuracy without overfitting target pronunciations. It implies the addition, selection and combination of relevant linguistic, phonological, prosodic and articulatory features. Then, finding the minimal quantity of material needed for training reliable models is of first importance because the cost of casting, segmenting and annotating speech databases is still very high. To provide robust conclusions, this question is studied on a variety of feature configurations.

In recent literature, models of pronunciation have been proposed for both automatic speech recognition (ASR) and TTS. Many statistical approaches have already been used for pronunciation modeling. Among them, neural networks [3–5] conditional random fields (CRFs) [3, 6, 7] and bayesian networks [8] are the most frequent. In the present work, pronunciation is modeled with CRFs. Only few studies report experiments on the quantity of speech training material [9, 10]. However, because the cost of data is still important it is necessary to evaluate data requirements in terms of size and content. Of course, statistics require large quantities of data, but in many fields of research –especially in affective computing– only small sized corpora are available, thus causing the problem of overfitting. In such cases, a compromise between both the quantity of training material and the size of the feature set needs to be reached. Whereas the search for data requirements has rarely been investigated, the search for an optimal feature set has been extensively studied, for example in the field of affective computing [11]. According to [12], with a small quantity of training material, reduced feature sets usually lead to models which better generalize than large feature sets. According to [13], "any subjective choice of which dimensions to keep and what heuristic reasoning to apply inevitably involves some assumptions about how the systems and workloads behave". As a consequence, a widely used method (also called brute-force method) begins with a large number of features, then performs dimension reduction. Another commonly used approach, set up

in the present work, is to introduce some human knowledge to select features *a priori*.

The following work improves the method proposed by [14] and adapts it to a French speech corpus. CRFs will be trained with different feature sets and different quantities of training data. These experiments are able to estimate which differences between phonemes generated by a phonetizer, and phonemes from the speech corpus, can be fixed up with a small speech corpus. Apart from improving TTS quality, the presented pronunciation adaptation method brings interesting perspectives in terms of expressive speech synthesis.

In the remainder, the speech corpus, its derived features and the experimental set-up are introduced in Section 2. Features and phoneme window selection experiments are presented in Section 3. Section 4 presents the training data reduction protocol and its results. A pronunciation example is discussed in Section 5. 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 used words in the telecommunication field. It features a neutral female voice sampled at 16 kHz (lossless encoding, one channel).

The corpus is composed of $7,208$ utterances, containing $196,190$ phonemes and $16,750$ non-speech sounds, totaling 5h49 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 [15].

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 consists in training 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, thereby leading to 52 feature set adapted from [14]. Most features have been normalized to corpus or utterance and discretized.

Canonical phonemes are generated with Liaphon [16], one of the most widely used utterance phonetization system for French. Word frequencies in French are extracted from Google n-grams [17]. 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 an oracle way, i.e., directly from the recorded utterances of the speech corpus. Such a protocol allows to know to 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 [18].

2.3 Experimental set-up

The phonemic sequences are modeled with CRFs, trained with the Wapiti toolkit [19]. Realized phoneme sequences and statistically adapted phoneme sequences are compared under the usual Phoneme Error Rate (PER). The speech corpus has been randomly split in two: training and development set (70%), and a validation set (30%). The training set has been divided in 7 folds. Models are trained on 6 folds, developed on 1 fold and tested on the remaining validation set. This protocol ensures that data used for training and testing do not overlap. The feature set at least consists in the canonical phoneme sequence generated with the phonetizer.

3 Optimal feature set

Finding an optimal feature set is a very important task in machine learning. It helps 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 for each group of features separately, using a forward selection process. Then groups of selected features are combined together with phoneme window to find the optimal configuration.

3.1 Feature selection within groups of features

Protocol A forward feature selection protocol has been adapted to French from previous work on spontaneous English pronunciation [14]. A cross-validation selection process was performed on the initial training set (six folds for training, one for testing) without any phoneme window. For each group of features, the selection starts with canonical phonemes only and other features are added one at a time until the optimal subset is reached. In order to find the global subset from the seven subsets obtained for each fold, a voting process has been set up.

Table 1: Selected features used for pronunciation modeling names and LPrPh feature set.

Group of feature	# feat.	Selected features
Linguistic (L)	2	Word ♦ Stem
Phonological (Ph)	7	Canonical syllables ♦ Syllable in word position ♦ Phoneme reverse position in syllable (numerical) ♦ Phoneme position and reverse position (numerical) ♦ Word length in phoneme (numerical) ♦ Pause per Syllable (low, normal, high)
Articulatory (A)	0	-
Prosodic (Pr)	6	Syllable Energy (low, normal, high) ♦ Syllable and phoneme tone (from 1 to 5) ♦ F_0 phoneme contour (decreasing, flat, increasing) ♦ Speech rate (low, normal, high) ♦ Distance to previous pause (from 1 to 3)

Selected features In the end, 15 linguistic, prosodic and phonological features were selected. Selected features are reported in the table 1. 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 [20], "noise reduction and consequently better class separation may be obtained by adding variables that are presumably redundant". Word expectation features, such as word frequency in French, received only very few votes. Surprisingly, it appears that no articulatory features have 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 them being related to phoneme positions in the utterance. None of the syllable characteristics (such as syllable part, structure or type) have been selected. Finally, six among seven prosodic features have been selected. This result is in agreement with state-of-the-art and suggests that a prosodic model is able to model a speaker's pronunciation.

3.2 Feature group combinations

Different combinations of selected feature groups were evaluated in cross-validation conditions, on the validation set without phoneme window. Average PER obtained on the seven folds are reported in bold in the table 2. The baseline is the PER obtained without any adaptation, between phoneme sequence generated by the phonetizer and realized phoneme sequence (ground truth). An improvement of 4.6 percentage point (pp) is reached while using a pronunciation

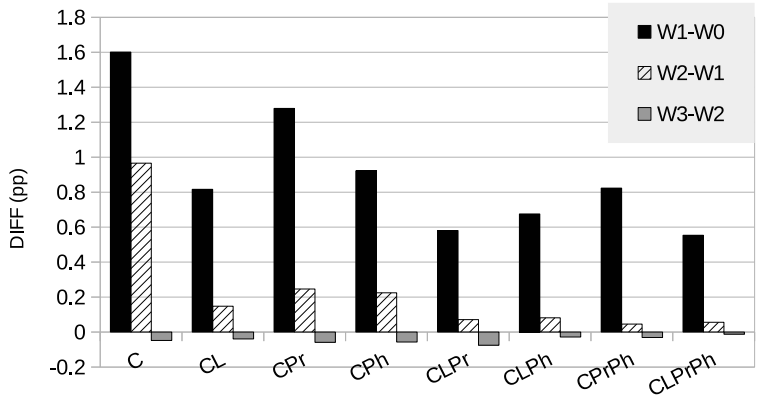

Fig. 1: Effect of the phoneme window size on the average PER obtained on 7 folds. $\text{DIFF}_i = \text{PER}(W_i) - \text{PER}(W_{i-1}), i \in \{1, 2, 3\}$.

model with canonical phonemes only, thus showing how pronunciation adaptation can reduce the inconsistency between the phonetized output and the speech corpus. Separately adding groups of selected features further improves the PER. The most spectacular reduction lies in the linguistic group: with only two apparently redundant features (word and its stem), a drop of 6.8 pp is obtained from the baseline. Overall results show an improvement in PER when combining selected feature groups. The combination of prosodic and linguistic groups lead to a significant drop in PER of 7.7 pp with a minimum number of features. The combination of the three feature groups brings the best PER, with an improvement of 7.9 pp from the baseline. In the end, only almost a third of the initial feature set remains.

3.3 Effect of phoneme window

In the search for an optimal feature set, a phoneme window is of real importance for pronunciation modeling since, linguistic, phonological and prosodic features of the current phoneme depend on the previous and next ones.

Protocol Only symmetrical windows have been tested since asymmetric windows did not show any interesting improvements [14]. The application of a symmetrical phoneme window Wx ($2 \cdot x + 1$ phonemes) is performed on the current canonical phoneme, and also its associated features therefore multiplying the number of features in the CRF model by x. Four phoneme window sizes are tested under cross-validation conditions using the same protocol as in the previous section. The results in terms of averaged PER on the seven folds and for different feature combinations are reported in the table 2. The relative gain ob-

tained while increasing the window size is represented on figure 1 for different feature combinations.

Table 2: PER values averaged on 7 folds in cross-validation conditions with adaptation. Different phoneme window sizes and different feature combinations. Baseline is 11.2%.

Window	C	CL	CPr	CPh	CLPr	CLPh	CPrPh	CLPrPh
W0	**6.6**	**4.4**	**4.8**	**4.5**	**3.5**	**4.0**	**3.7**	**3.3**
W1	5.0	3.6	3.6	3.6	2.9	3.3	2.8	2.8
W2	4.1	3.4	3.3	3.3	2.9	3.2	2.8	2.7
W3	4.1	3.5	3.4	3.4	3.0	3.2	2.8	2.7

Results Figure 1 shows that adding features coming from one (black) or two (hatched) surrounding neighbours have a positive effect on the global PER. A seven phoneme window (W3) degrades the results, probably because as the number of feature increases, the model overfits the data. The effect of the phoneme window size differs according to the combination of feature used for training models. For instance, a phoneme window has a higher effect when models are trained with prosodic features than linguistic or phonological features. Indeed, prosodic features of the current phoneme highly depends on what precedes and what succeeds. Finally, according to table 2, the combination of a window W2 and the 15 selected features brings the best results. In the next section, four configurations are tested: two phoneme window W0 and W2, and two feature sets: canonical phonemes only (C) and the 15 selected features with canonical phonemes (CLPrPh).

Perceptive tests were realized on synthesized speech samples with different feature combinations [21]. As in [22], the results are strongly linked with PER and confirmed the relevancy of both the pronunciation adaptation model and the selected features. Some samples are available on the team website[1].

4 Minimal training data size

Because the cost of pronunciation corpora is very high, it is worth trying to find the minimal quantity of training material required for pronunciation adaptation. The obtained results would tell us the expected accuracy for a given duration of training material. For a given quantity of training material, models are evaluated in terms of PER in cross-validation conditions.

[1] http://www-expression.irisa.fr/demos/: Corpus-specific adaptation

4.1 Protocol

The training set (70% of the initial speech corpus) has been divided in $N_f = 7$ folds: 6 for training and the remaining for development purpose. The different size of training material is obtained while splitting the initial training set in 2×7, then 4×7, 8×7, etc. At each step, 6 folds are used for training models, one of the remaining folds is kept for development. In order to limit the experimental time, we have limited N_f to 100 for training durations less than 300 min While the quantity of training material decreases, N_f increases thus making results more reliable: from 243.3 min of training data ($N_f = 7$, 4321 utterances each) to 40 s of training data ($N_f = 100$, 12 utterances each). The validation set consists in 120.2 min of data and 2161 utterances. Two feature sets (C and CLPrPh) and two phoneme windows (W0 and W2) are tested. This choice allows to study the effects of the number of features on accuracy and to estimate the danger of using too much features while training on too few data.

4.2 Results

As expected, the averaged results on figure 2 (top) show that the smaller the duration of training data, the higher the phoneme error rate. What is surprising is that models trained with very small data improve the PER of 4.0 pp. from baseline (best configuration W0-CLPrPh). Therefore, small training sets allow fixing many phoneme errors: recurrent pronunciation, alphabet mapping, French schwa and liaison (see section 5). Of course, with very small training set, standard deviation computed on all the folds increase. Therefore, it appears that some sets allow to reach a good PER, whereas some others do not. Thus, the minimal quantity of training material does not lie in its duration only, but also in the content of this set.

PER as a function of the duration of training data follows two different trends whether duration is over 4.4 min or not. Interestingly, for training duration over this threshold, the logarithmic curve of PER is almost linear with respect to duration (correlation coefficient over 0.96, see table 3). This result is in agreement with the results obtained in ASR experiments [10]. For small durations (less than 4.4 min), phoneme window has almost no effect on PER whereas the number of feature does. When training models with very little data, there is a weak effect of window and feature sets (PER range is 0.9 pp.). The results show that CRF models trained with small datasets are still better than the baseline. All the more, multiplying the duration by 6.6 leads to an improvement of 2.6 pp. (best W0-CLPrPh configuration). For larger durations of training data (more than 4.4 min), feature set has less effect compared to phoneme window. Interestingly, increasing the amount of training data does not improve significantly the accuracy of models trained with W0-C configuration. In this case, multiplying the duration by 10 leads to an improvement of only -0.5 pp. (best W2-CLPrPh configuration).

The obtained results show that there is a threshold for duration of training data at almost 5 min. Over this threshold, the addition of new data has a high

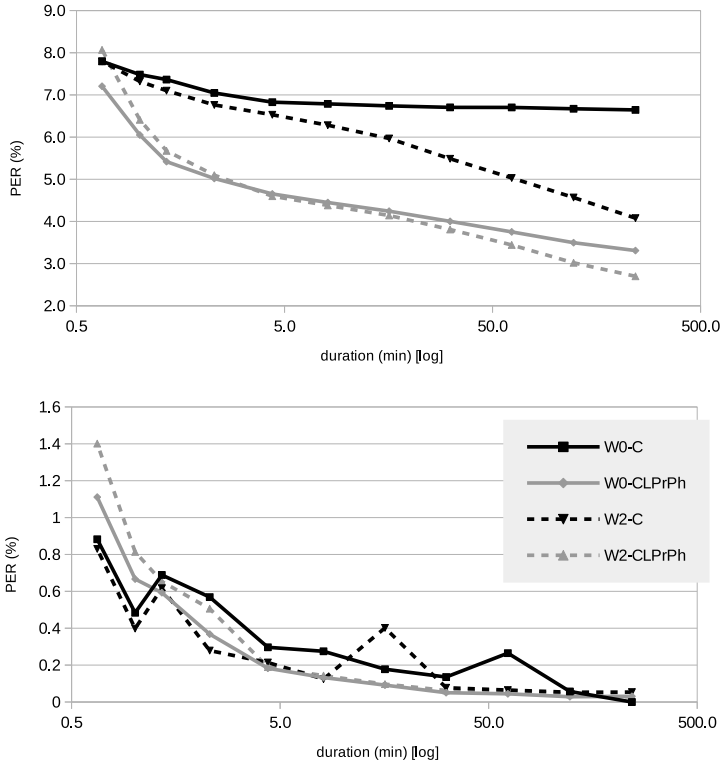

Fig. 2: Average (top) and standard deviation (bottom) PER between canonical and adapted phonemes obtained on the validation set. Average is computed on all available folds for a given training duration (log-scale minutes). Baseline is 11.2 %.

Table 3: Linear regression results of PER w.r.t. training data duration (logarithmic scale).

Training duration	Lin. Reg.	W0-C	W0-CLPrPh	W2-C	W2-CLPrPh
> 0.7 min	Slope	-0.17	-0.54	-0.58	-0.73
	Corr. coef.	0.74	0.85	0.99	0.86
> 4.0 min	Slope	-0.04	-0.34	-0.62	-0.48
	Corr. coef.	0.96	1.00	0.99	0.99

cost but a weak improvement in accuracy. Since the PER is log-linear with respect to the duration, an ideal PER $= 0$ would be reached for $3 \cdot 10^8$ hours of training data with W2-CLPrPh configuration.

5 Discussion

Table 4: Example of pronunciation adaptations with different windows, features and training size. The input text is *Dans la montagne, les couleurs sont exceptionnelles.* "In the mountains, colors are remarkable"

Win.	Features	dur(min)	Phoneme sequence
Realized			d ɑ̃ l a m ɔ̃ t a **n j** - l e k u l œ ʁ s ɔ̃ **t** ɛ k s ɛ p s j **o** n ɛ l -
Canonical			d ɑ̃ l a m ɔ̃ t a **ɲ** - ə l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j ɔ n ɛ l ə
W2	CLPrPh	243.3	d ɑ̃ l a m ɔ̃ t a **n j** - l e k u l œ ʁ s ɔ̃ z ɛ k s e p s j ɔ n ɛ l -
W2	C	243.3	d ɑ̃ l a m ɔ̃ t a **n j** - l e k u l œ ʁ s ɔ̃ t e k s ɛ p s j **o** n ɛ l -
W0	C	243.3	d ɑ̃ l a m ɔ̃ t a **n j** - l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j ɔ n ɛ l -
W2	CLPrPh	4.4	d ɑ̃ l a m ɔ̃ t a **n j ə** l e k u l œ ʁ s ɔ̃ **t** ɛ k s ɛ p s j **o** n ɛ l -
W2	C	4.4	d ɑ̃ l a m ɔ̃ t a **n j** - l e k u l œ ʁ s ɔ̃ **t** ɛ k s ɛ p s j **o** n ɛ l -
W0	C	4.4	d ɑ̃ l a m ɔ̃ t a **n j** - l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j **o** n ɛ l -
W2	CLPrPh	0.7	d ɑ̃ l a m ɔ̃ t a **g** - **e** l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j **o** n ɛ l -
W2	C	0.7	d ɑ̃ l a m ɔ̃ t a **ʁ** - - l e k u l œ ʁ s ɔ̃ **t** ɛ k s ɛ p s j **o** n ɛ l -
W0	C	0.7	d ɑ̃ l a m ɔ̃ t a **ʁ** - - l e k u l œ ʁ s ɔ̃ - ɛ k s ɛ p s j ɔ n ɛ l -

CRF models trained with small datasets bring better results than the baseline in terms of PER, hence underlying the power of such pronunciation adaptation models. Models trained with very few utterances are able to fix some regular errors between canonical and realized phonemes: recurrent pronunciation, French schwa and liaisons. Generally, canonical phonemes and realized phonemes are not encoded using the same alphabet, therefore introducing phoneme differences which are not typical errors. Interestingly, CRF are able to solve the alphabet issues.

Table 4 shows an example of adaptation results on the pronunciation of an utterance in the validation set. This example illustrates typical errors. First, a one-phoneme window (W0) is not able to model French liaisons (in the example: /s ɔ̃ **t** ɛ/), whatever the duration of the training set. A larger phoneme window combined with C or CLPrPh is able to model the liaison, but the result is not always correct (/z/ instead of /t/ with 243.3 min of training data). CRF models trained with 40 s of data are not able to label correctly the canonical symbol /ɲ/: labels /n j/ are not found but /ʁ/ or /g/. The deletion of French schwa is realized in all configurations at the end of the utterance. Models trained with the full CLPrPh feature set and few data label the schwa with either /e/ or /ə/ (probably because models overfit the data). The substitution /ɔ/ → /o/ is better modeled with a large phoneme window. CRF models are able to map alphabets from canonical to realized. For example, the symbol /ɲ/ in the canonical sequence does not exist in the alphabet used for the realized phoneme annotation. Most models trained with more than 4 min of data are able to adapt the canonical symbol to the realized one. In the context of speech synthesis, some phoneme

errors are more harmful than others. For example in Table 4, the substitution /o/ → /ɔ/ or /o/ is not as important as the substitution of /n,j/ → /ʁ/ or /g/. Therefore an adapted error rate based on how close are phonemes could improve statistical approaches for speech synthesis.

6 Conclusion

In this paper, we have presented a pronunciation adaptation method which adapts phonemes generated by the phonetizer to the speech corpus. A CRF pronunciation model trained with linguistic, phonological, articulatory and prosodic features predicts an adapted phoneme sequence from a canonical phoneme sequence. The present work investigates an optimal feature set (features and phoneme window) and a minimal quantity of training material for TTS.

First a cross-validation forward feature selection methodology is proposed. This method allows to select 15 linguistic, phonological and prosodic features (LPrPh). Different feature group combinations are tested together with different phoneme window sizes. An optimal feature set (W2-CLPrPh) brings the best improvement (-8.5 pp) in terms of PER on the validation set. Hence, we have shown that pronunciation adaptation to the speech corpus itself helps to significantly reduce the inconsistency between phonemes as labeled in the underlying speech corpus and those generated by the phonetizer. Moreover, a statistical approach has the advantage of being easily reproducible.

Second, a cross-validation experiment was conducted with decreasing quantities of training material. We can conclude from these experiments that there is a threshold for duration of training data at almost 5 min. Over this threshold, the addition of new data has a high cost but a weak improvement in accuracy: multiplying the duration of training data by 10 improves the PER of 0.5 pp. An ideal PER = 0 would be reached for $3 \cdot 10^8$ hours of training data with W2-CLPrPh configuration. Therefore for exploratory researches on pronunciation, 5 min of training data seem to be enough. However, for end-user applications, the more data, the better.

The advantage of large-scaled data is to introduce several weighted phoneme adaptations for each canonical phoneme. Then, including n-best predicted phonemes into phoneme lattices together with weighted phoneme errors could be relevant for speech synthesis applications. Apart from improving TTS quality, the presented pronunciation adaptation method brings interesting perspectives in the use of small-scaled speech corpora for expressive TTS.

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

References

1. Olinsky, C., Cummins, F.: Iterative English adaptation in a speech synthesis system. In IEEE Workshop on Speech Synthesis (2002)

2. Govind, D., Prasanna, S.M.: Expressive speech synthesis: a review. International Journal of Speech Technology, vol. 16, 237–260 (2013)
3. 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)
4. Rao, K., Peng, F., Sak, H., Beaufays, F.: Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In Proc. of ICASSP (2015)
5. Yao, K. and Zweig, G.: Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. In Proc. of Interspeech (2015)
6. Lecorvé, G., Lolive, D.: Adaptive Statistical Utterance Phonetization for French. In Proc. of ICASSP (2015)
7. Hazen, T.J., Hetherington, I., Shu, H., Livescu, K.: Pronunciation modeling using a finite-state transducer representation. Speech Communication, vol. 46, 189–203 (2005)
8. Livescu, K., Jyothi, P., Fosler-Lussier, E.: Articulatory feature-based pronunciation modeling. Computer Speech and Language, vol. 36, 212–232 (2016)
9. Nagòrski, A.,Boves, L., Steeneken, H.: In search of optimal data selection for training of automatic speech recognition systems. In Proc. of ASRU (2003)
10. Moore, R. K.: A comparison of the data requirements of automatic speech recognition systems and human listeners. In Proc. of Eurospeech (2003)
11. Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V.: The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In Proc. of Interspeech (2007)
12. Tahon, M., Devillers, L.: Towards a small set of robust acoustic features for emotion recognition: challenges. IEEE/ACM Transaction in Speech, Audio and Language Processing, vol. 54, Issue 1, 16–48 (2016)
13. Chen, Y., Ganapathi, A., Katz, R.: Challenges and opportunities for managing data systems using statistical models. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering (2011)
14. Qader, R., Lecorvé, G., Lolive, D., Sébillot, P.: Probabilistic speaker pronunciation adaptation for spontaneous speech synthesis using linguistic features. In Proc. of SLSP (2015)
15. Chevelu, J., Lecorvé, G., Lolive, D.: ROOTS: a toolkit for easy, fast and consistent processing of large sequential annotated data collections. In Proc. of (LREC) (2014)
16. Béchet, F.: LIA-PHON: un système complet de phonétisation de texte. Traitement Automatique des Langues (TAL), vol. 42, 47–67 (2001)
17. Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., Petrov, S.: Syntactic annotations for the Google books ngram corpus. In Proc. of ACL (2012)
18. d'Alessandro, C., Rosset, S., Rossi, J.-P.: The pitch of short-duration fundamental frequency glissandos. Journal of Acoustical Society of America, vol. 104, 2339–2348 (1998)
19. Lavergne, T., Cappé, O., Yvon, F.: Practical very large scale CRFs. In Proc. of ACL (2010)
20. Guyon, I., Elissef, A.: An introduction to variable and feature selection. Journal of Machine Learning Research, vol. 3, 1157–1182 (2003)
21. Tahon, M., Qader, R., Lecorvé, G., Lolive, D.: Improving TTS with corpus-specific pronunciation adaptation. In Proc. of Interspeech (2016)
22. Qader, R., Lecorvé, G., Lolive, D., Sébillot, P.: Adaptation de la prononciation pour la synthèse de la parole spontanée en utilisant des informations linguistiques. In Proc. of Journées d'Etudes sur la Parole (2016)