Error detection of grapheme-to-phoneme conversion in text-to-speech synthesis using speech signal and lexical context

Kévin Vythelingum, Yannick Estève, Olivier Rosec

ERROR DETECTION OF GRAPHEME-TO-PHONEME CONVERSION IN TEXT-TO-SPEECH SYNTHESIS USING SPEECH SIGNAL AND LEXICAL CONTEXT

Kévin Vythelingum[1,2], Yannick Estève[2], Olivier Rosec[1]

[1] Voxygen, Pleumeur-Bodou, France
[2] LIUM, Le Mans University, France

ABSTRACT

In unit selection text-to-speech synthesis, voice creation involved a phonemic transcription of read speech. This is produced by an automatic grapheme-to-phoneme conversion of the text read, followed by a manual correction. Although grapheme-to-phoneme conversion makes few errors, the manual correction is time consuming as every generated phoneme should be checked. We propose a method to automatically detect grapheme-to-phoneme conversion errors by comparing contrastives phonemisation hypothesis. A lattice-based forced alignment system is implemented, allowing for signal-dependent phonemisation. We implement also a sequence-to-sequence neural network model to obtain a context-dependent grapheme-to-phoneme conversion. On a French dataset, we show that we can detect to 86.3% of the errors made by a commercial grapheme-to-phoneme system. Moreover, the amount of data annotated as erroneous is kept under 10% of the total evaluation data. The time spent for phoneme manual checking can thus been drastically reduced without decreasing significantly the phonemic transcription quality.

Index Terms— automatic error detection, grapheme-to-phoneme conversion, forced alignment, sequence-to-sequence neural networks, speech synthesis

1. INTRODUCTION

Text-to-speech synthesis (TTS) consists in generating a speech signal from an input text. There are several paradigms for TTS, including unit selection speech synthesis [1] and statistical parametric speech synthesis [2]. Recently, works have been done to replace some or all components of traditional TTS systems by neural networks. We can cite Wavenet [3], Tacotron [4] and Deep Voice [5, 6]. However, most commercial systems are still based on unit selection TTS, where speech is generated by the concatenation of acoustic units selected in a speech corpus. This technique permits high quality TTS due to the human nature of the speech signal. In order to create a synthetic voice for a unit selection TTS system, a voice talent read aloud a dedicated text. All the recorded utterances are phonetically segmented: in general, phonemic transcriptions are derived from text with grapheme-to-phoneme conversion (G2P) systems and aligned automatically on the speech signal.

G2P consists in converting a sequence of words into a sequence of phonemes. Numerous approaches were proposed in the literature to derive automatically pronunciation from words. The most popular are dictionary look-up, rule-based systems [7] and joint n-gram models [8, 9]. Besides, G2P can be considered as a machine translation task, where the problem is to translate a sequence of characters into a sequence of phonemes [10]. More recently, state-of-the-art results were reached on standard English G2P tasks with sequence-to-sequence models inspired by neural machine translation [11, 12].

However, G2P systems still make serious errors and the annotation of speech databases should be very accurate, as the TTS quality highly depend on the phonetic segmentation accuracy. In [13], the authors showed that manually corrected phonemic transcriptions in French TTS datasets can improve speech synthesis quality. Moreover, [14] gives evidences that a better phoneme accuracy benefits to synthesis. Thus, a manual checking is necessary to correct the remaining errors of the G2P system. This task is very time consuming as it requires to inspect all the generated phonemes to check if they are all representative of what the speaker really said. We focus on reducing the amount of phonemes we need to check manually when an automatic G2P system is used to derive phonemic sequences from text. In other words, we want to detect the errors of G2P to help manual correction of TTS datasets. For this purpose, we compute signal-dependent phonemic transcriptions and G2P hypothesis from a context-dependent neural network model. The error detection relies on the consensus between the different systems. To the best of our knowledge, our proposed method is the first attempt to detect G2P errors using signal-dependent phonemic transcriptions and context-dependent neural network hypothesis.

The paper is organized as follows. In section 2, we present the architecture of our error detection system, which relies on a signal-dependent phonemic transcription and a neural model trained on already corrected phonemic transcriptions. It includes also the description of the G2P system we want to detect the errors, which is based on a set of rules and a morpho-syntactic analysis. Then, the results on French datasets used for commercial TTS are discussed in section 3, followed by a conclusion.

2. ERROR DETECTION SYSTEM COMPONENTS

A G2P system infers phonemic sequences from text. However, it is sometimes impossible to choose the right pronunciation of a word only with its spelling. That's why context-dependent G2P is used to take decisions with additional knowledge such as part-of-speech. Another way to disambiguate word pronunciations is to exploit the speech signal as it is available in TTS datasets. A TTS dataset is indeed formed by an ensemble of read-speech, word-level transcription, phone-level transcription and segmentation. For this purpose, we built an acoustic model to obtain a signal-dependent phonemic labeling with forced alignment. The error detection relies on the comparison of the resulting signal-based phonemic transcription and the text-based G2P conversion. When some data is manually corrected, we can use it to train a context-dependent data-driven G2P model. We show finally that we can have a better error detection using this system to produce an additional phonemic hypothesis.

As shown in Figure 1, the error detection system consists in four major components:

- The **rule-based G2P system** converts from written text to phonemes. It is the system we want to detect the errors.
- The **data-driven G2P system** produces a context-dependent phonemic hypothesis.
- The **forced alignement system** produces a signal-based phonemic transcription.
- The **comparison module** aligns phonemic sequences and put *correct* and *error* labels for respectively matching and mismatching phonemes. When more than two inputs are compared, the *correct* label is put only if all phonemes are identical.

Text is fed to the rule-based G2P system to generate phonemes. Then, these phonemes are manually corrected to obtain the reference. The two resulting phonemic sequences are finally compared to put *correct* and *error* labels for each phoneme of the G2P hypothesis. This gives the error detection reference.

Besides, we process forced alignment of audio and text to generate another phonemic sequence hypothesis. It consists in using an acoustic model to align the pronunciation provided by a lexicon on the speech signal. This phonemic hypothesis is then compared to the rule-based G2P hypothesis to obtain the error detection hypothesis, which is then compared to the error detection reference for evaluation.

In addition, we train a data-driven grapheme-to-phoneme on manually corrected phonemic transcriptions. It gives an additional G2P hypothesis for error detection based on a different use of the lexical context than the rule-based G2P system. We implement indeed a character-based sequence-to-sequence neural model.

2.1. Rule-based Grapheme-to-phoneme System

The rule-based G2P system is a proprietary system used for French TTS. It is composed by three modules: a lexicon, a set of transliteration rules and a morphosyntactic analyser. The lexicon gives the pronunciation of words and the transliteration rules are used as a fallback for words that are not present in the lexicon. As several pronunciations can be possible for the same word, the G2P system disambiguates the different hypotheses according to the part-of-speech of words given by a morphosyntactic analyser.

2.2. Data-driven Grapheme-to-phoneme System

In addition to the rule-based G2P system, we develop a G2P conversion system based on sequence-to-sequence neural network modeling with an attention mechanism. It takes character-level word transcriptions as input and outputs phone-level phonemic transcription. The model benefits from the manually corrected data and fits particularly for context-dependent phonemic transcription. Indeed, our model is based on the encoder-decoder architecture developed in [15], with the exception that the decoder is composed by two gated recurrent unit (GRU) layers interleaved with attention mechanism, the hidden state of the decoder is initialized with a non-linear transformation applied to the mean bi-directional encoder state and the maxout hidden layer before the softmax operation is removed. In fact, we follow the configuration of the default attention model of the open-source nmtpy toolkit [16], with 64-dimensional embeddings and 128-dimensional hidden layers. During training, we use dropout with probability 0.4 after each recurrent layer. We also use the Adam optimization algorithm with a batch size of 32 and a learning rate of 10^{-4}.

2.3. Forced alignment system

The forced alignment system involves two components: an acoustic model and a lexicon.

Firstly, we trained a GMM-HMM acoustic model on perceptual linear prediction (PLP) features with feature space maximum likelihood linear regression (fMLLR) speaker adaptation. Then, we trained a DNN-HMM model using the frame-level cross entropy criterion based on the fMLLR speaker adapted PLP features and the senone alignment from the GMM-HMM model.

The DNN model follows this topology :

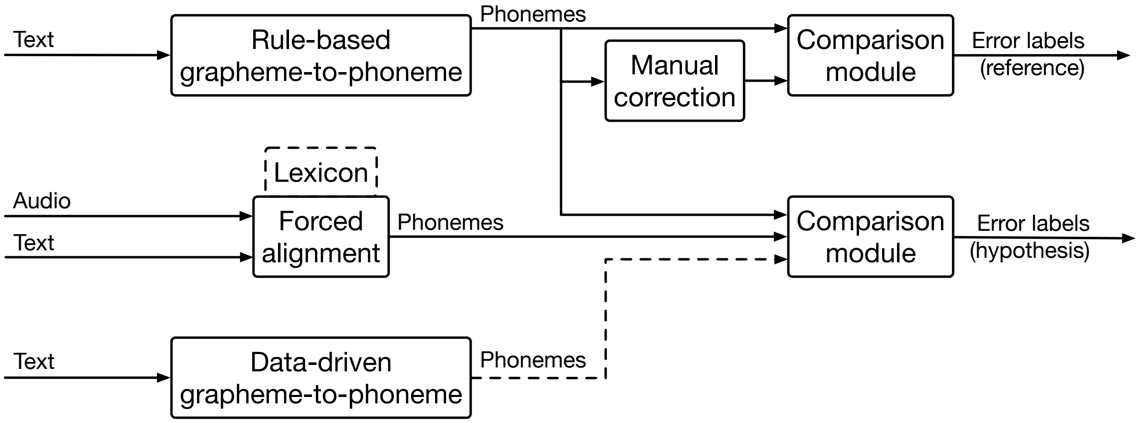

Fig. 1. System diagram depicting error detection procedure, with inputs on the left and outputs on the right.

- an input layer of 360 dimensions (the input features of 40 dimensions are spliced across 8 neighboring frames)
- five hidden layers of 3000 dimensions
- an output layer of 10553 dimensions

These acoustic models were trained using the Kaldi speech recognition toolkit [17].

The lexicon is first built by applying the rule-based G2P system on the list of all words of the dataset. As the G2P conversion is processed without lexical context, several pronunciation hypothesis for each word are given. However, some pronunciations alternatives may still miss for some words. That is why we enriched the lexicon by adding hypothesis from a statistical G2P conversion model. We used the Phonetisaurus toolkit [18, 19] to build an alignment model and a n-gram based translation model implemented as a weighted finite state transducer (WFST). For the translation model, we computed a 6-gram language model based on the grapheme-phoneme alignment using SRILM [20, 21]. Then, we processed the forced alignment with several choices for the number of additional pronunciation hypothesis in the enriched lexicons.

2.4. Comparison module

The comparison module aligns phonemic sequences using the NIST SCLITE tool in order to put *correct* and *error* labels for respectively matching and mismatching phonemes. To take into account the hypothesis of the data-driven G2P system in addition to the forced alignment phonemic transcription, we also combine the outputs of the error detection of each system, putting the *correct* label only if both systems gives a *correct* label when compared to the rule-based G2P hypothesis.

3. RESULTS

We train our models using internal French TTS datasets containing approximately 50 hours of speech data from 9 speakers segmented into 90,135 utterances. The results are then given by testing our models on internal French TTS datasets containing approximately 10 hours of speech data from 3 speakers segmented into 16,328 utterances.

3.1. Grapheme-to-phoneme results

The evaluation data contains 16,328 segments, 125,433 words and 427,768 phonemes. The data-driven G2P system is trained on a character-level grapheme-phoneme bitext with a symbol for word separation. Table 1 gives a representation example of the bitext corpus. The corpus contains 90,135 segments, 618,155 words and 2,120,794 phonemes, with a vocabulary of 33,191 words.

Table 1. Representation example of the bitext corpus

Graphemes	l e s | é c r a n s | s o n t | a l l u m é s
Phonemes	L EI | Z EI K R AN | S ON | T A L U M EI

Table 2. Phone error rates of grapheme-to-phoneme systems

	PER (%)
rule-based G2P system	1.8
data-driven G2P system	1.4

The performance of the rule-based G2P system and the data-driven G2P system is given in terms of phone error rate

(PER), which is the mean percentage deviation in Levenshtein distance with the manually corrected phonemic transcription. As show in table 2, the data-driven G2P system obtains a slightly better PER than the rule-based system.

3.2. Error detection results

The evaluation data contains 427,768 phonemes, of which 1.8% are erroneous. Error detection performance are measured with three evaluation metrics. Precision and Recall are standard metrics to evaluate error detection systems. They indicate respectively the proportion of true alarms raised by the error detection system and the proportion of detected errors. We introduce also the Manual Checking Rate (MCR), which shows the amount of data which is annotated as erroneous by the system. We want to maximize Precision and Recall while we want to minimize MCR.

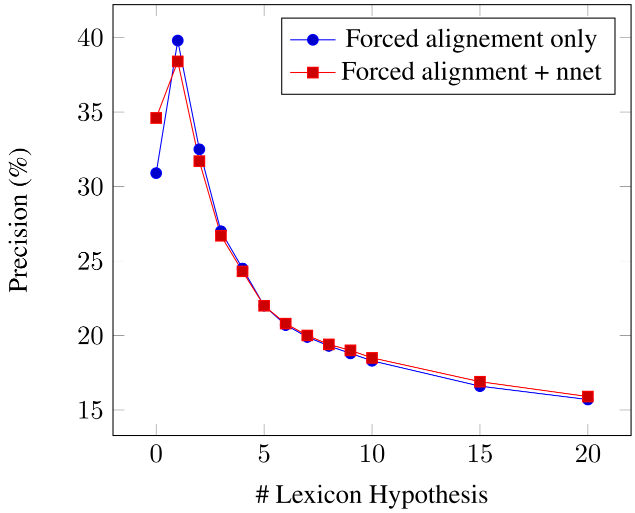

Fig. 2. Precision of the error detection system according to the number of statistical hypothesis added to the forced alignment lexicon.

Figure 2 shows the Precision of the error detection system according to the number of statistical hypothesis added to the forced alignment lexicon. Except when we use the baseline lexicon without any enrichment, the combination of forced alignment and neural based phonemisation have the same Precision. When we add one statistical hypothesis per word from the joint n-gram G2P model, the Precision increases drastically from 30.9% to 39.8%. Then, the Precision decreases quickly to 20.7% when we enrich the lexicon with 5 more pronunciations per word, and decline slowly to 15.7% for 20 hypothesis. A compromise with the gain in Recall should be found to obtain a reasonable amount of data to check manually.

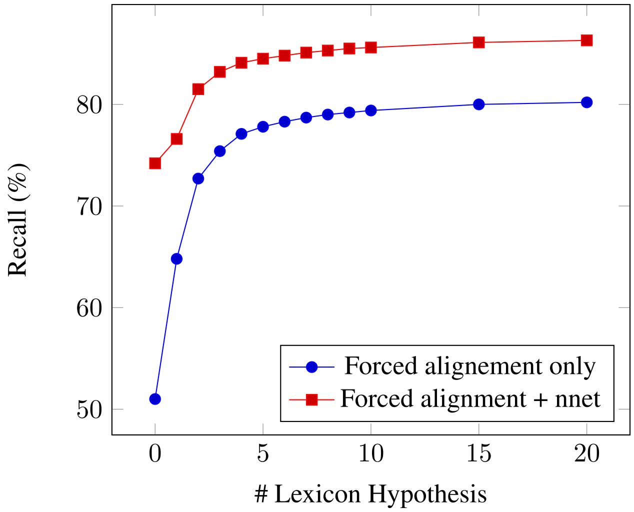

Fig. 3. Recall of the error detection system according to the number of statistical hypothesis added to the forced alignment lexicon.

Figure 3 shows the Recall of the error detection system according to the number of statistical hypothesis added to the forced alignment lexicon. The Recall increases by augmenting the number of statistical hypotheses added in the lexicon used for forced alignment. For forced alignement only, a asymptote of 80% is reached. This corresponds to the case when we add the pronunciations variants from the reference to the baseline lexicon. However, the combination of forced alignment and neural based G2P pushes this limit to more than 85%. It is clear that combining signal based and neural based phonemic transcription helps to detect more errors than using only forced alignment.

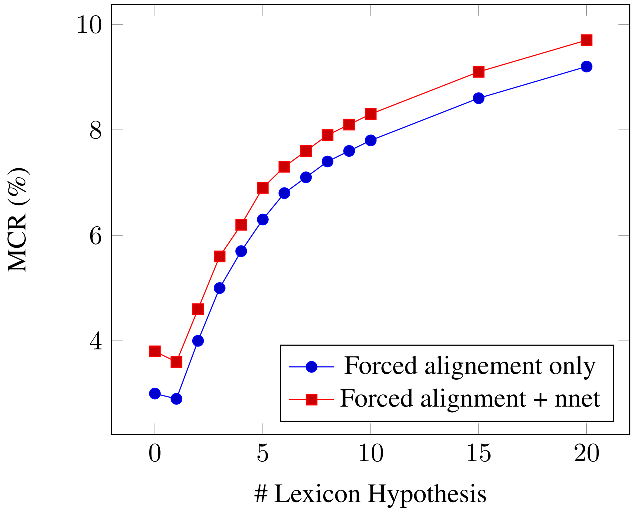

Fig. 4. Manual Checking Rate of the error detection system according to the number of statistical hypothesis added to the forced alignment lexicon.

Fig. 4 shows the MCR of the error detection system according to the number of statistical hypothesis added to the forced alignment lexicon. The two curbs follow the same trend but forced alignment combined with neural based G2P involved slightly more MCR than forced alignment only. As the MCR continues to increase relatively fast comparing to the improvements for Recall, it seems reasonable to not add more than 4 statistical phonemisations to the forced alignment lexicon. This gives respectively for forced alignment only and forced alignment combined with neural based G2P a MCR of 5.7% and 6.2%, and a Recall of 77.1% and 84.1%. However, it is possible to reach respectively a Recall of 80.2% and 86.3% with a MCR under 10%. In other words, a human annotator can correct up to 86.3% of G2P errors by checking less than 9.7% of the dataset.

4. CONCLUSION

We proposed a method for error detection of grapheme-to-phoneme conversion in text-to-speech synthesis. Our approach takes advantage of the audio available in speech synthesis datasets. By using forced alignment with an acoustic model, we obtained a contrastive phonemic transcription in comparison to the one we want to correct. The error detection is then improved by enriching the forced alignment lexicon with statistical G2P hypothesis. Experimental results show that this can help manual correction of phonemic transcriptions in TTS datasets, which is a critical task for commercial TTS. As we noticed, with our approach a human annotator can correct up to 86.3% of G2P errors by checking less than 9.7% of the data. We show also that a neural based G2P trained on already corrected datasets improve error detection when combined with the forced alignment system. Further work will consist in increasing the precision of error detection and validating our approach with other languages than French.

5. REFERENCES

[1] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, pp. 373–376.

[2] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” in *Speech Communication*, 2009, pp. 1039–1064.

[3] A. Van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” in *arXiv:1609.03499v2*, 2016.

[4] Y. Wang, R.J. Skerry-Ryan, D. Stanton, Y. Wu, R.J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, “Tacotron: Towards end-to-end speech synthesis,” in *arXiv:1703.10135*, 2017.

[5] S. Ark, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, “Deep voice: Real-time neural text-to-speech,” in *arXiv:1702.07825v2*, 2017.

[6] S. Ark, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Zhou Y., “Deep voice 2: Multi-speaker neural text-to-speech,” in *arXiv:1705.08947v1*, 2017.

[7] F. Béchet, “Lia phon : un système complet de phonétisation de textes,” in *Traitement Automatique des Langues (TAL)*, 2001, pp. 47–67.

[8] M. Bisani and H. Ney, “Joint-sequence models for grapheme-to-phoneme conversion,” in *Speech Communication*, 2008, pp. 434–451.

[9] L. Galescu and J. F. Allen, “Pronunciation of proper names with a joint n-gram model for bi-directional grapheme-to-phoneme conversion,” in *Proceedings of InterSpeech*, 2002.

[10] A. Laurent, P. Delglise, and S. Meignier, “Grapheme to phoneme conversion using an smt system,” in *Proceedings of InterSpeech*, 2009.

[11] K. Rao, F. Peng, H. Sak, and F. Beaufays, “Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4225–4229.

[12] K. Yao and G. Zweig, “Sequence-to-sequence neural net models for grapheme-to-phoneme conversion,” in *Proceedings of InterSpeech*, 2015.

[13] S. Brognaux, Picart B., T. Drugman, and Louvain D., “Speech synthesis in various communicative situations: Impact of pronunciation variations,” in *Proceedings of InterSpeech*, 2014.

[14] R. Dall, S. Brognaux, K. Richmond, C. Valentini-Botinhao, G.E. Henter, J. Hirschberg, Yamagishi J., and S. King, “Testing the consistency assumption: Pronunciation variant forced alignment in read and spontaneous speech synthesis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, p. 51555159.

[15] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate.,” in *International Conference on Learning Representations (ICLR)*, 2015.

[16] O. Caglayan, M. García-Martínez, A. Bardet, W. Aransa, F. Bougares, and L. Barrault, “Nmtpy: A flexible toolkit for advanced neural machine translation systems,” in *arXiv:1706.00457*, 2017.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burge, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The kaldi speech recognition toolkit,” in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[18] J. R. Novak, P. R. Dixon, N. Minematsu, K. Hirose, C. Hori, and H. Kashioka, “Improving wfst-based g2p conversion with alignment constraints and rnnlm n-best rescoring,” in *Proceedings of InterSpeech*, 2012.

[19] J. R. Novak, N. Minematu, and K. Hirose, “Failure transitions for joint n-gram models and g2p conversion,” in *Proceedings of InterSpeech*, 2013.

[20] A. Stolcke, “Srilm – an extensible language modeling toolkit,” in *Proceedings of InterSpeech*, 2002.

[21] A. Stolcke, J. Zheng, and Wen Wang, “Srilm at sixteen: Update and outlook,” in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.