INTERSPEECH 2009 BRIGHTON

Improving Consistence of Phonetic Transcription for Text-to-Speech

Pablo Daniel Agüero[1], *Antonio Bonafonte*[2], *Juan Carlos Tulli*[1]

[1]Communications Lab, University of Mar del Plata, Argentina
[2]TALP Research Center, Universitat Politècnica de Catalunya, Spain
pdaguero@fi.mdp.edu.ar

Abstract

Grapheme-to-phoneme conversion is an important step in speech segmentation and synthesis. Many approaches are proposed in the literature to perform appropriate transcriptions: CART, FST, HMM, etc. In this paper we propose the use of an automatic algorithm that uses the transformation-based error-driven learning to match the phonetic transcription with the speaker's dialect and style. Different transcriptions based on word, part-of-speech tags, weak forms and phonotactic rules are validated. The experimental results show an improvement in the transcription using an objective measure. The articulation MOS score is also improved, as most of the changes in phonetic transcription affect coarticulation effects.

Index Terms: speech synthesis, speech intelligibility, speech recognition, Hidden Markov models

10.21437/Interspeech.2009-194

1. Introduction

In text-to-speech synthesis (TTS) an important part of the effort to build a new voice is spent on preparing the database. As a consequence, it is crucial to reduce the amount of effort needed in building it. Manual supervision increases the database preprocessing costs. However, it is not a warranty of success and the final result may have manual errors or inconsistencies introduced by different points of view of reviewers.

Grapheme-to-phoneme (g2p) conversion is a very important part of the text-processing module of text-to-speech synthesis systems, both for database segmentation and speech synthesis. In the literature we found several methods based on machine learning (and not just lexicons): CART [1], FST [2], HMM [3], Transformation-based error driven learning [4], and a comparative evaluation of other methods can be found in Damper et al [5]. Those approaches focus on several aspects of automatic phonetic transcription, such as the treatment of unknown words (missing in the lexicon), pronunciation variants, and the transcription of names (for example: proper names, foreign names, etc.).

The phonetic transcription is used for automatic phone segmentation to prepare a speech database for speech synthesis. In general it is assumed that a correct transcription is known, for example, by means of a lexicon plus a manual correction.

The problem of automatic phone segmentation may be solved with three different approaches depending on the previous information we may have: unconstrained, acoustically constrained or linguistically constrained [6].

Although some researchers claim that actual automatic segmentation systems can already achieve accurate enough results for their use on speech synthesis [7], many research labs still get their best results by manually supervising the data. Therefore, in the literature we can still find many research works on a variety of methods such as: Hidden Markov Models [7, 8], Artificial Neural Networks [9] or Dynamic Time Warping [10, 11].

Errors in the phonetic transcription introduces noise in the estimation of statistical parameters of phone models, such as HMM. Insertions, deletions or incorrect phonetic transcription induces the system to include in the estimation wrong information. What's more, an incorrect transcription may cause concatenation problems, because the phone is not placed where it is expected.

Another important problem that arises are the systematic errors in phonetic transcription as a sign of dialectal differences between the lexicon and the speaker. A dialect is a variety of a language that is characteristic of a particular group of the language's speakers. In general, a dialect corresponds to regional speech patterns, but a dialect may also be defined by other factors, such as social class (sociolect). Due to the number of speakers and the geographical area covered by them, the dialect can be of arbitrary size and might contain several sub-dialects. Therefore, it is normal that once the dialect of the speaker is chosen, some deviations from standard may appear due to many reasons: geographical, social, cultural, etc.

Our experience in Blizzard 2008 [12] showed us the difficulty to find an appropriate phonetic transcription for the speakers in UK English and Mandarin Chinese. The transcription provided by the UPC's text-to-speech system named Ogmios did not accurately match the style of the speakers. As a consequence, identity and intelligibility scores in the subjective evaluation were lower than expected.

In this paper we propose an automatic method to validate a set of transcriptions in a lexicon and a set of manually written rules. The goal is to obtain an optimal transcription for database segmentation and speech synthesis that preserves speaker's style. The quality measure used in the approach is the phone log-likelihood of HMM models. The automatic algorithm uses the transformation-based error-driven learning algorithm proposed by Brill [13].

This paper is organized as follows. In Section 2, the proposed algorithm is explained in detail. In Section 3, the results of the experiments are presented and discussed. Finally, the main conclusions are summarized in Section 4.

2. Proposed algorithm

The proposed algorithm in this paper will be compared with the baseline grapheme-to-phoneme conversion applied by Ogmios for UK English in the evaluation of Blizzard 2008 [14]. It consists of three steps, as shown in Figure 1.

The first step of the baseline system is the phonetic transcription of the words using a lexicon according to the orthographic transcription of the words and their part-of-speech tags (POS). The lexicon consists of a list of words and the corre-

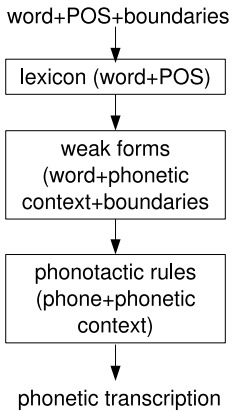

Figure 1: Grapheme-to-phoneme conversion process in Ogmios for UK English used in Blizzard 2008.

sponding phonetic transcription for each possible POS of the word. If the word has several transcriptions with the same POS, the most frequent is considered. The lexicon is built (or chosen) for the same dialect spoken by the speaker.

The second step is the transcription of words which have weak forms. In English, most words will have at least one stressed syllable, and hence no separate strong and weak forms. All words which do have distinct strong and weak forms are monosyllables, and are usually function words or discourse particles. For most of these, the weak form is the one usually encountered in speech. As the extreme example, the strong form of the indefinite article *a* is used only in the rare cases when the word is stressed: naming the word, or when emphasizing indefiniteness.

The main words with weak forms are: a, am, an, and, are, as, at, be, been, but, can, could, do, does, for, from, had, has, have, he, her, him, his, just, me, must, of, shall, she, should, some, than, that, the, them, there, to, us, was, we, were, who, would and you.

The third step in the baseline grapheme-to-phoneme conversion in Ogmios is the use of a set of phonotactic rules to adjust the phonetic transcription according to language constraints. Phonotactics is a branch of phonology that deals with restrictions in a language on the permissible combinations of phonemes. It defines permissible syllable structures, consonant clusters, and vowel sequences by means of phonotactical constraints.

The main problem of the baseline approach is the rigid structure of rules of transcription that do not have a validation according to speaker's style. The speaker may have small deviations from the canonical transcription and they should be considered to improve both speech segmentation and text-to-speech synthesis. The immediate consequence of those mismatches in the canonical transcription may be a lower score in speaker's identity and intelligibility.

As explained in the introduction, the speakers of a given dialect may have deviations from the canonical transcription due to many reasons, such as geographical, social, cultural, etc. Therefore, it is necessary to find these particularities to improve the quality of automatic segmentation and the identity of the synthesized speech through the speaker's style.

In this paper we propose to study the validness of the transcription rules using the phone log-likelihood of HMM using forced alignment.

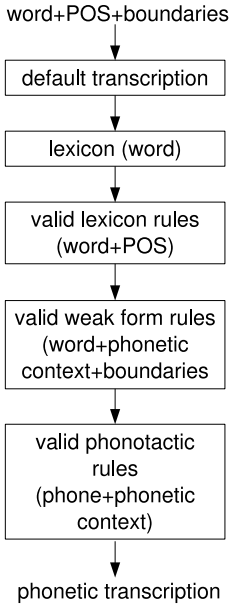

Figure 2: Proposed grapheme-to-phoneme conversion process in Ogmios for UK English to improve performance obtained in Blizzard 2008.

Figure 2 shows the proposed phonetic transcription scheme with rule validation. It has the same steps that Figure 1, with two additional steps: default transcription and lexicon (word). The default transcription of the word is chosen given the most probable part-of-speech, provided by the part-of-speech tagger. It is identical to the transcription of the first step in the baseline system.

The second step of our proposal includes all possible transcriptions without considering the POS. In this way we obtain an optimal phonetic transcription that is independent of the POS tag. The other steps are identical to the baseline system. However, each rule in those steps will be validated using the training data.

The validness of each rule is studied in an incremental way. Each new rule is used to perform a new forced alignment of all the speech database. The difference between the previous global log-likelihood (without applying the new rule) and the present global log-likelihood (including the new rule) is defined as the gain of the rule.

Those rules with a positive gain are kept as valid rules and the process continues with the following rules under evaluation. The rules are sequentially studied in each step: lexicon(word), lexicon rules (word+POS), weak form rules (word+phonetic context+boundaries) and phonotactic rules (phone+phonetic context). After each step HMM models are retrained to adjust the HMM parameters to the improved phonetic transcription.

The weak forms provide different transcription depending on the context, such as the presence of a pause before the word, or the type of the first phone of the following word: vowel or consonant. The processing order of the rules may cause different transcriptions due to the variation of the phonetic context. This effect is not relevant because the phonemes do not change their type: vowel or consonant.

Phonotactical rules introduce modifications in the phonetic transcription of the words based on the position of the phonemes

in the word, their identity and the phonetic context. An error in the phonetic transcription does not only affect the word, but also the adjacent words. For example, the elision of a phoneme may force an alignment of an existing phoneme (deleted in the transcription by a rule) into the adjacent phones, decreasing their log-likelihood. As a consequence, the score is calculated for all the phonemes and not only those phonemes affected by the rule.

A common approach in speech database building is the use of alternative word transcriptions to improve the quality of speech segmentation. The goal of our proposal is to improve both speech segmentation and text-to-speech synthesis. The consistency between the rules applied for both processes is necessary to improve intelligibility and to preserve the speaker's identity due to peculiarities of his/her dialect.

3. Experiments

The experiments in this paper were performed using the subset ARCTIC of the 15 hours UK English database released by the Centre for Speech Technology Research (CSTR) for Blizzard 2008.

The lexicon is based on the Unisyn dictionary provided by the University of Edinburgh [15]. It consists of 117K word entries. After listening to some samples, the accent chosen for this task is the RP accent. SAMPA was selected as the phoneset.

The forced-alignment of phones is performed by the HMM-based ASR named Ramses [16]. The HMM models are trained using context-independent models. A silence model, trained at punctuation marks, was optionally inserted at each word boundary during the alignment to enable pause detection.

Previous experiments support the automatic segmentation approach. They have shown that when a correct phonetic transcription is given, HMM models can achieve similar speech synthesis quality than manual segmentation [7, 17]. In addition, Adell et al. [18] showed how the log-likelihood in Hidden Markov models may be used to eliminate phonemes with wrong transcription in the speech database.

The number of rules at each step of the validation process in the experiment performed with 1,098 utterances were:

- **lexicon (word)**: 208 rules.
- **lexicon (word+POS)**: 57 rules.
- **weak form rules**: 52 rules.
- **phonotactic rules**: 9 rules.

The first step to validate lexicon rules without considering part-of-speech tags gave an improvement in the log-likelihood for 49 word transcriptions. In the 208 rules under study 9 words have one alternate transcription, 95 have two possible alternate transcriptions and 3 have three transcriptions. An extract of the first twelve valid rules detected in the first step is show in Table 1.

The next step validated lexicon rules using part-of-speech tags. The rules affected the following words: *a, again, as, at, bowed, but, conduct, do, either, grave, had, has, have, her, moderate, of, or, them* and *to*.

The validated weak forms rules were the corresponding to the following words after a pause: *a, and, as, be, can, does, from, has, of, shall, into, unto, were* and *will*. All of them got the stressed form. The rules that reduce the words *are* and *the* before a consonant were also validated.

Finally, the only valid phonotactic rule was the elision of */r/* at the end of the word when followed by other consonants.

Word	SAMPA transcription
a	@
again	@ - g 'eI n
against	@ - g 'eI n s t
and	'{ n d
an	@ n
articulate	A: - t 'I - k j u - l @ t
as	@ z
at	@ t
because	b I - k @ z
bow	b 'aU
bowed	b 'aU d
but	b @ t

Table 1: Valid rules at first step

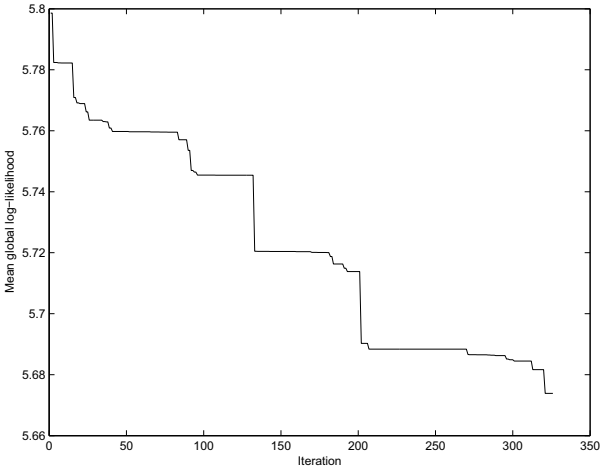

Figure 3: Improvement in each validation step.

In Figure 3 we show the evolution of the mean global log-likelihood when evaluating each rule. There is a monotonic improvement of the log-likelihood during all rule validation process. At iterations 208, 265 and 317 is shown the improvement due to retraining of HMM models.

The highest improvement in the mean global log-likelihood at step one corresponds to words *a* (iteration 3), *and* (iteration 15), *an* (iteration 17), *had* (iteration 91), *of* (iteration 132) and *was* (iteration 201). Another relevant improvement is found at iteration 321 and corresponds to the only valid phonotactic rule at step four.

The performance of the new grapheme-to-phoneme conversion module using validated rules was evaluated using a subjective test on three different aspects: similarity (S) of the synthesized voice to the real speaker, quality (Q) of the synthesis and the naturalness and correctness of the articulation(A). Eight native British English speakers performed the perceptual test using a five point scale (1:worst and 5:best).

The box-plots of Figure 4 do not show any relevant differences in similarity and quality between the baseline and the proposed system (SB vs SP and QB vs QP, respectively). The proposed approach did not improve the quality of the baseline system due to the limitation on the amount of data used in the experiments.

The main improvement in the subjective results is shown in

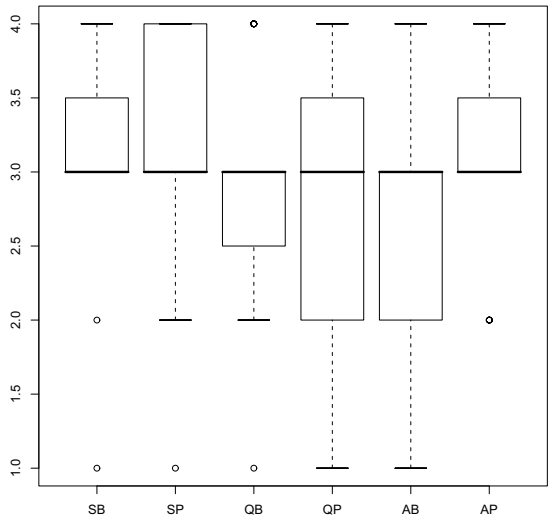

Figure 4: Similarity, quality and articulation Mean Opinion Scores.

the articulation MOS (AB vs AP). The evaluators preferred the articulation of the proposed approach rather than the baseline system. The Wilcoxon test shows that the difference is relevant for $p < 0.01$.

Although the general quality is the same for the baseline system and the new one, the later is perceived as better articulated. In fact, as most of the changes in phonetic transcription affect coarticulation effects, the results are reasonable.

Notice that the database used in our experiments is quite formal (read style). We expect that the method will be even more relevant if applied to non-professional speakers or to less formal data.

4. Conclusions

Automatic grapheme-to-phoneme conversion presents some limitations due to small deviations from canonical transcription in our text-to-speech system named Ogmios. They are originated in the dialect and style of the speaker.

In this paper we propose a technique to validate a set of rules at different steps of automatic phonetic transcription: lexicon, lexicon+POS, weak forms and phonotactic. The proposed procedure checks the validness of the rules using a sequential approach.

The experiments show a set of valid rules at the different steps extracted using the proposed approach. The total gain in the mean global phoneme log-likelihood is 2.15% compared to the baseline system presented at Blizzard 2008. This better phonetic transcription has a direct impact on the quality of segmentation and text-to-speech synthesis, as shown in the MOS of articulation.

Although the preliminary subjective results shown in this paper are encouraging, we will continue working with new experiments on other databases and a higher amount of evaluators to obtain more statistically relevant conclusions.

5. References

[1] Hunt, A. and Black, A., "Unit selection in a concatenative speech synthesis system using a large speech database", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1:373–376, 1996.

[2] Galescu, L. and Allen, J., "Bi-directional conversion between graphemes and phonemes using a joint n-gram model", Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.

[3] Taylor, P., "Hidden markov models for grapheme to phoneme conversion", Proceedings of Interspeech, 1973–1976, 2005.

[4] Polyakova, T. and Bonafonte, A., "Using error-driven approach to improve automatic grapheme-to-phoneme conversion accuracy", TC-STAR Workshop on Speech-to-Speech Translation, 213–217, 2006.

[5] Damper, R.I., Marchand, Y., Adamson, M.J. and Gustafson, K., "Comparative evaluation of letter-to-sound conversion techniques for English text-to-speech synthesis", Proceedings of 3rd ESCA/COCOSDA International Workshop on Speech Synthesis, 53–58, 1998.

[6] Marzal, A. and Vidal, E., "A review and new approaches for automatic segmentation of speech signals", Proceedings EUSIPCO, Barcelona, 43–53, 1990.

[7] Makashay, M. J., Wightman, C. W., Syrdal, A. K. and Conkie A., "Perceptual evaluation of automatic segmentation in Text-to-Speech synthesis", Proceedings of the International Conference on Spoken Language Processing, 431–434, 2000.

[8] Taylor, P. A. and Isard, S. D., "Automatic phone segmentation", Proceedings of Eurospeech, 709–711, 1991.

[9] Toledano, D. T., Hernandez Gomez, A. and VIllarrubia Grande, L., "Automatic phone segmentation", IEEE Transactions on Speech and Audio Processing, 2(6):617–625, 2003.

[10] Malfrère, F. and Dutoit, T., "High quality speech synthesis for phonetic speech segmentation", Proceedings of the European Conference On Speech Communication and Technology, 2631–2634, 1997.

[11] Kominek, J., Bennet, C. and Black, A. W., "Evaluating and correcting phoneme segmentation for unit selection synthesis", Proceedings of Eurospeech, 313–316, 2003.

[12] Black, A. W., King, S. and Tokuda, K., "The Blizzard Challenge 2008 - Evaluating corpus based speech synthesis on common databases", Blizzard Challenge, 2008.

[13] Brill, E., "Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging", Computational linguistics, vol. 4:543–565, 1995.

[14] Bonafonte, A., Moreno, A., Adell, J., Agüero, P. D., Banos, E., Erro, D., Esquerra, I., Perez, J. and Polyakova, T., "The UPC TTS system description for the 2008 Blizzard Challenge", Blizzard Challenge, 2008.

[15] Fitt, S., "Documentation and user guide to unisyn lexicon and post-lexical rules", 2000.

[16] Bonafonte, A., Mario, J. B., Nogueiras, A. and Rodriguez Fonollosa, J. A., "RAMSES: el sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC", Proceedings of VIII Jornadas de Telecom I+D (TELECOM I+D'98), 399–408, 21998.

[17] Adell, J., Bonafonte, A., Gomez, J. A., and Castro, M. J., "Comparative study of automatic phone segmentation methods for TTS", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1: 309–312, 2005.

[18] Adell J., Agüero, P. D. and Bonafonte, A., "Database pruning for unsupervised building of text-to-speech voices", Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 1:889–892, 2006.