Fusion of dictionaries in voice creation and speech synthesis task

Tatyana Polyakova, Antonio Bonafonte
Universitat Politècnica de Catalunya,
tatyana@gps.tsc.upc.edu, antonio.bonafonte@upc.edu

Abstract

The accurate phonetic transcription is very important for different fields of speech technologies. In speech synthesis, it is the benchmark for the voice segmentation, and therefore one of the crucial points for the synthesized speech pronunciation quality. In ASR the availability of matching phonetic transcription allows a higher recognition precision. Use of different dictionaries could improve the phonetic transcription since it allows a better word coverage, but the "direct" dictionary merging presents incompatibility problems. Dictionary fusion is the method that automatically learns the dictionary-to-dictionary transformation rules. The results presented in this paper show that fusion significantly improves the compatibility between the dictionaries (about 32-83% of improvement), and allows reducing the number of "severe" pronunciation errors in comparison with the grapheme-to-phoneme (g2p) techniques.

1. Introduction

Nowadays, the demand for high quality speech technologies has risen significantly for all fields of application. TTS and ASR systems gain thousands of the new users daily. The quality of actual non-commercial systems is good but still some improvements are highly desirable. In many ways the quality of these systems depends on the accuracy of the phonetic transcription. Normally, the phonetic transcription module is based on a system dictionary and an automatically trained g2p converter is used for the derivation of the pronunciation of the unknown word. The average automatic g2p conversion results for English can vary between 70 and 85 percents of words correct, depending on the method and the dictionary used for the evaluation [1, 3, 5].

Recently the UPC's speech synthesis research group has participated in two international speech synthesis evaluation campaigns, whose goal was to compare the TTS systems in the framework of the TC-STAR project and Blizzard challenge.

In order to have more extensive dictionary coverage the available dictionaries were merged. The merging was motivated by the idea that it was better to have the transcription from other dictionary than not to have any previously validated transcription. Only the words that didn't appear in the system dictionary were added from the auxiliary dictionary. The phone sets of all the dictionaries in question were compared and homogenized, in other words the phonemes that were not present in the added dictionaries were mapped to the closest corresponding phonemes from the reference dictionary. The term reference dictionary designates the dictionary that was chosen by an expert to be matching the speaker's dialect. Even after making all these changes we are not sure that the criteria used by different experts to produce the dictionaries were the same.

The goal of this paper is the detailed study the inconsistencies and incompatibility issues between the dictionaries and find out if a higher level of compatibility could be reached by applying automatic rules, and if dictionary fusion can improve g2p results.

2. Dictionary description

This work studies three different dictionaries CMU dictionary, LC-STAR dictionary and UNISYN dictionary.

A dictionary is a list of words with the corresponding phonetic transcriptions. Some dictionaries, in addition, have morphosyntactic information and the syllable boundaries, but no additional information was used in this work.

The CMU dictionary, provided by Carnegie Mellon University, includes about 125K North American words which were generated using independent sources of proper and common names among which were expert proofed transcriptions as well as some synthesizer-generated ones. The phonetic transcription had to be converted to Sampa.

The Unisyn dictionary, provided by the University of Edinburgh (www.cstr.ed.ac.uk) consists of 110K word entries, all of which are common English names. The great advantage of this dictionary is that it is transcribed in metaphonemes which allows the encoding of multiple accents of English (UK, US, Australian and New Zealand accents are included). Output is available in Sampa or IPA. The accent chosen for this task is the rhotic version of the NYC accent.

The LC-STAR dictionary was created in the framework of the LC-STAR project (www.lc-star.org). The dictionary includes 50K common and 50K proper North American names. Each proper name is marked with a label whether it is a geographic, person's or company name. LC-STAR project created dictionaries in 13 languages.

In this work LC-STAR was chosen to be the target dictionary because it had both proper and common names labeled. Since the proper names present a special problem they had to be studied separately. The unlabeled proper names from CMU dictionary were found using as the guide the labeled proper names from the LC-STAR dictionary.

It is desirable to have the transcription criteria to be the same because, more coherent phonetic transcriptions allow to build better models in ASR and to better define speech segments in TTS.

The first step was to discover which one of the two source dictionaries was closer to the target dictionary. To validate the "direct" dictionary merging method previously used in the TTS evaluations, the words that appeared in both dictionaries of each pair were compared (each dictionary was compared with LC-STAR). The comparison was performed directly, by selecting all the common entries for each pair of the dictionaries and counting the errors. The phone sets were adapted to be the same. For example, the phoneme /4/ that represents a voiced alveolar tap is absent from the CMU dictionary so it was replaced by the phoneme /t/. The rhotic /@r r/ and /3r r/ of the Unisyn was replaced by /@r/ and /3r/ correspondingly.

Some examples of differences between dictionaries are: k @ m j u t **@** d vs. k @ m j u t **I** d; h ae l @ p i n **j o** vs. h ae l @ p i n **o;** t **A O r** m i n @ vs. t **aU @r** m i n @**;** m @ g E r i vs. m @ **k** g E r i, k j **3r** vs. k j **u r,** etc. The comparison results are given in the Tables 1 and 2.

Table 1. The comparison results for CMU and LC-STAR dictionaries (common and proper names)

CMU vs. LCSTAR common		CMU vs. LCSTAR proper	
with stress	without stress	with stress	without stress
phonemes 93.37%	phonemes 94.17%	phonemes 86.95 %	phonemes 88.03%
words 69.10 %	words 71.06%	words 57.23%	words 58.28%

Table 2. The comparison results for Unisyn and LC-STAR dictionaries (common names)

UNISYN vs. LCSTAR common	
with stress	without stress
phonemes 84.14 %	phoneme 84.86 %
words 31.44 %	words 32.02%

Objectively, from tables given above we can conclude that there is a higher level of similarity between the common names of CMU and LC-STAR dictionaries, than between Unisyn and LC-STAR. CMU seems to be rather similar but still far from reaching 100% compatibility. Assuming both of these transcriptions are correct makes the validity of the g2p module evaluations using a reference transcription from the evaluator's dictionary arguable. The pronunciation of proper names seems to be quite different between the dictionaries. The Unisyn dictionary seems unsuitable for

merging with LC-STAR dictionary before performing the pronunciation adaptation. The difference between the stress patterns for all dictionary pairs is unimportant, but seems to be greater for proper names. From now on only the results on the dictionaries without stress will be considered.

3. Dictionary Fusion in TTS

One of the goals of dictionary fusion is to use automatically adapted dictionary to improve the g2p results and therefore the overall speech quality. The whole procedure can be represented by the scheme in Fig 1.

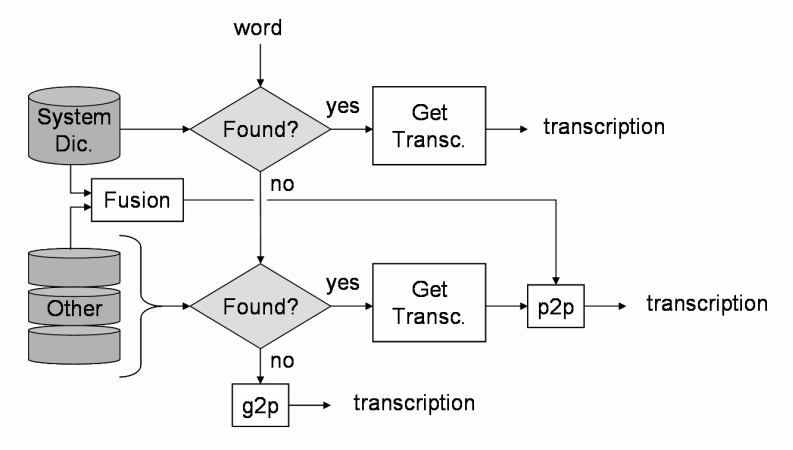

Fig 1. Scheme of the phonetic module in our TTS system

First, the input word it searched in the system dictionary, if it is found the Get Transc. tool outputs the corresponding pronunciation. If the input is not found in the system dictionary, then it is searched in the auxiliary dictionary and if it is present there, the fusion is performed. The transcription is obtained form the auxiliary dictionary and then passed to the p2p (phone-to-phoneme) module, which adapts the words pronunciation to be compatible with the system dictionary. In the last case, if the word is not found in any of the dictionaries, the transcription is obtained by passing a grapheme string to the g2p module.

4. Fusion method and g2p techniques

In this section we are going to explain the g2p and p2p techniques. In the first case, given the orthographic form (word) we want to obtain the phonetic form (pronunciation), while in the p2p case we take the word found in both of the dictionaries in question and learn the mapping between the source and target phoneme strings. Two machine learning techniques decision tress (DT) [1] and finite-state transducers (FST) [4] were used for training both the automatic g2p and p2p conversion systems. These methods require an alignment between the source and target dictionaries.

In order to perform p2p conversion between dictionaries, given that the length difference between source and target strings are unpredictable, it was necessary to have an alignment method able to introduce the nulls both into source and target strings, that is why the dynamic programming algorithm based alignment method was implemented [3].

Since there is no way to predict where to put nulls in the source string, these were removed and the phonemes corresponding to the null letters were joined by an underscore with the previous phoneme, e.g. [b] [o] [x] [_] / [b] [A] [k] [s] -> [b] [o] [x] / [b] [A] [k_s]

For the g2p case, in order to have more consistency in the alignments, the list of prohibited alignments was used. No vowel-consonant, consonant-vowel alignments were allowed. The

description of the machine-learning methods, used in all of the experiments, is given in the next two sub sections.

4.1 Decision Trees (DT)

A classical machine learning technique often applied to obtain grapheme-to-phoneme transcriptions is the method based on decision trees [1]. This method is appropriate for discreet characteristics and produces rather compact models, whose size is defined by the total number of questions and leaf nodes in the output tree. In our case we applied it both to g2p and p2p conversion.

A decision tree has as the input source characters sliding window with three characters to the left and three to the right of each center character accordingly.

4.2 Finite State Transducers (FST)

This technique [4] chooses the pronunciation $\hat{\varphi}$ that maximizes the probability of a target sequence given the source sequence g. This is equivalent to:

$$\hat{\varphi} = \arg\max_{\varphi} \{ p(\varphi / g) \} = \arg\max_{\varphi} \{ p(g, \varphi) \} \quad (1)$$

Estimate the probability (1) is the same as maximize the probability of the source-target pair. This estimation can be done using standard n-gram methods. Source-target pairs are extracted from the aligned dictionary.

$$p(g,\varphi) = \prod_{i=1}^{N} p(g_i, \varphi_i / g_1^{i-1}, \varphi_1^{i-1}) \quad (2)$$

where N is the number of characters in the source [2]. The FST was also applied both to p2p and g2p conversion.

5. Experimental results

The research done in the framework of this paper set several goals. The first one was to compare dictionaries transcribed by independent producers and to see if the criteria used were very different. The second objective was to find out whether it was possible to, at least partially, eliminate the inconsistencies existing in these dictionaries, if considered from the point of view of the dictionary that matches the speaker's accent, automatically. The last goal was to see if it was reasonable to use the automatically adapted dictionary to improve the g2p results and therefore the overall speech quality.

For the evaluation of the p2p converter we considered the common entries from each pair of dictionaries as in Tables 1 and 2. The training set for each experiment consisted in 90% percent of the common entries and the test set of 10% accordingly.

The fusion results for 3 different pairs of dictionaries are given below. Table 3 shows the results for the conversion of common and proper names from CMU to LC-STAR format, Table 4 from Unisyn to LC-STAR respectively.

Table3. Common and proper names conversion CMU to LCSTAR dictionary

CMU => LCSTAR common		CMU => LCSTAR proper	
DT	FST	DT	FST
phonemes 97.26%	phonemes 96.80%	phonemes 88.91%	phonemes 88.40%
words 84.72%	words 83.70%	words 58.30%	words 59.26%

Table 4. Common names conversion Unisyn to LC-STAR dictionary

UNISYN => LCSTAR common	
DT	FST
phonemes 96.36%	phonemes 96.72%
words 79.92%	words 83.02%

To justify the importance of the dictionary fusion the g2p conversion was performed both for common and proper names form the LC-STAR dictionary. The g2p conversion results are shown in the Table 5.

Table 5. G2p results for common and proper names for LC-STAR dictionary

LCSTAR g2p common		LSTAR g2p proper	
DT	FST	DT	FST
phonemes 94.22%	phonemes 96.11%	phonemes 87.34%	phonemes 89.91%
words 70.09%	words 81.58%	words 53.10%	words 65.42%

The test set was the same as for the p2p conversion for CMU=>LC-STAR scheme, which is 10% percent of the common words between those dictionaries, while the training set in each of the cases included all the remaining words in the corresponding dictionary.

From the Tables 3 and 4 we can conclude that the fusion is able to reduce the number of inconsistencies existing between dictionaries, especially for common names, where the after-fusion word accuracy improvements range about 13 % for the CMU dictionary and about 50 % for the Unisyn dictionary. After the fusion the latter one reaches a similar level of compatibility with the reference that CMU. Nevertheless, there are still some different transcription criteria that the automatic methods were not able to capture.

The p2p accuracy for the fusion of CMU and LC-STAR dictionaries is higher than the g2p accuracy for the LC-STAR dictionary, therefore allowing us to think that the use of the dictionary fusion as a part of phonetic transcription module could improve the speech quality. Furthermore, we believed that g2p conversion errors could be more severe than differences between dictionaries, for this reason, fifty erroneous words for 3 conversion schemes (cmu->lcstar , unisyn->lcstar , g2p) were analyzed and "quickly" classified into 3 categories, according to the criteria from Table 6. The 3 categories were: L, M and S. The "light" errors "L", do not difficult the comprehension of the word, may even pass unnoticed to a non-native English speaker. The medium, "M", errors could make the word less recognizable but not unpleasant to hear, while the "S" errors can severely affect the word's comprehensibility when pronounced by a synthesis system, and can sound unpleasantly.

Table 6. Classification of errors, examples and criteria.

Type	Criteria used	Example
L	• shwa substituted by other short vowel • flap /4/ substituted by /t/ or vice versa • short vowel substituted by a similar long one and vice versa • missing shwa • a short vowel substituted by shwa • consonant-consonant confusion (within same category)	@ -> I 4 -> t, t -> 4 I->i, i-> I A-> @, I -> @, E -> @ s-> z, t>d
M	• ¨two or three errors of type "L" • missing consonant • affricate->fricative, fricative-plosive etc. confusions • diphthong-vowel confusions	 dZ->z, Z->z ae ->aI , aI -> i
S	• two or more errors of type "M" • vowel-consonant confusions • more than three errors of type "L"	3r->r, A -> r, V -> v

Table 7. Count of erroneous examples for each conversion method for each category

	L	M	S
cmu -> lcstar	38	10	2
unisyn->lcstar	39	8	3
g2p	28	15	7

From the Tables 3,4,5,7 we can observe that in the case of common names by fusing the dictionaries we do not only obtain a higher overall pronunciation accuracy but also reduce the number of severe and moderate errors that really worsen the speech quality.

6. Conclusions

After carrying out all of the above described experiments, several conclusions can be made. The experiments confirm the existence of significant number of inconsistencies between different dictionaries; some of them appear due to the difference in transcription criteria employed by the experts while others are caused by the inconsistencies already existing in the dictionaries which, in its turn could be a result of using various unhomogenized sources for building the dictionary (like in the case if CMU dictionary). Some of these differences can be overcome by dictionary fusion procedure which consists deriving automatic p2p conversion rules from the words that the dictionaries have in common. In the case of the common words form LC-STAR dictionary it seems to be feasible to fuse them only with those CMU dictionary since this procedure gives the best results. For proper names the fusion does not give significant improvements due to the elevated difficulty of the proper names transcription problem even for human experts. The DT and FST methods give similar results in the fusion task. After performing fusion the number of severe transcription errors decreases, therefore guaranteeing a better quality of synthesized speech.

7. Acknowledgements

The authors would like to thank Carnegie Mellon University and the University of Edinburgh for providing the CMU and Unisyn dictionaries respectively. This work was partially funded by TC-STAR and AVIVAVOZ projects.

References

1. Black A.W., Lenzo K. and Pagel V., “Issues in building general letter to sound rules”, In Proceedings of the Third ESCA workshop on speech synthesis, Jenolah Caves, W-S W, Australia, 1998

2. Bonafonte A. and J. B. Mariño, “Language modeling using X-grams”, In Proc. of ICSLP-96, Vol.1, Philadelphia, pp. 394-397, 1996

3. Damper R. I., Marchand Y., Marsterns J.-D. and Bazin A., “Aligning letters and phonemes for speech synthesis” in Proceedings of the 5th ISCA Speech Syntesis Workshop, Pittsburgh, 209-214., 2004

4. Galescu L., J. Allen, “Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model”, In Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland, 2001

5. Polyakova T., Bonafonte A., "Learning from errors in grapheme-to-phoneme conversion", International Conference on Spoken Language Processing, Pittsburgh, USA, 2006.