DATABASE PRUNING FOR UNSUPERVISED BUILDING OF TEXT-TO-SPEECH VOICES

Jordi Adell, Pablo Daniel Agüero, Antonio Bonafonte

Dept. of Signal Theory and Comunications
TALP Research Center
Universitat Politècnica de Catalunya
Barcelona, Spain
www.talp.upc.edu

ABSTRACT

Unit Selection speech synthesis techniques lead the speech synthesis state of the art. Automatic segmentation of databases is necessary in order to build new voices. They may contain errors and segmentation processes may introduce some more. Quality systems require a significant effort to find and correct these segmentation errors. Phonetic transcription is crucial and is one of the manually supervised tasks. The possibility to automatically remove incorrectly transcribed units from the inventory will help to make the process more automatic. Here we present a new technique based on speech recognition confidence measures that reaches to remove 90% of incorrectly transcribed units from a database. The cost for it is loosing only a 10% of correctly transcribed units.

1. INTRODUCTION

Unit selection speech synthesis is reaching the highest performance nowadays. While other type of synthesisers do not generate as much intelligible and natural speech, unit selection is the state of art [1, 2, 3]. However, such technique works the best when a big database is available. It requires for a large inventory of units in different contexts. Then it is necessary to record databases, typically hours, and transform them into an inventory of units that can be successfully managed by the synthesis module.

In order to create such inventory from scratch, i.e. from audio files and their corresponding prompt text, it is necessary to normalise the text, phonetically transcribe it and segment recorded speech into phoneme or diphone units. For such big databases it is either very expensive to manually perform these tasks or problems in automatic processes would generate undesired units. These undesired units may be due to misplaced boundaries or to incorrectly transcribed units.

Furthermore, new voices are often requested for several applications. For example, in order to create emotional synthesis [4, 5]. Since the synthesiser speaking style strongly depends on the recorded speech, every time a new style is requested a new voice has to be recorded. Also for synthesisers in new languages the process to create a new voice is very expensive. Therefore, making the process more automatic will be very helpful.

Since containing wrong units in the inventory of a speech synthesis system can cause spotting errors, it is desirable to detect them before building its inventory. Segmentation when transcription is known is a solved problem as previous experiments have shown [6, 7], then the main problem relies on correctly transcribe the units that have been said. The second is mainly a problem of speech decoding, however here we are not as interested in correctly transcribe all the units as in detecting problematic ones. We assume, as it has previously been done in other related works [8, 9], that wrong units will not be a big portion of the database and that it is affordable to loose such part of it. Therefore we focus on detecting undesired units in order to be able to remove them from the inventory.

In Section 2 we describe the complete system we use to build unit inventories. Then, in Section 3 methods proposed to detect incorrectly transcribed units are explained. Afterwards, in Section 4 and 5 the evaluation and experiments performed are described. Finally, conclusions are discussed.

2. OVERALL SYSTEM OVERVIEW

In this section we describe the method we are using in order to automatically build voices for our unit selection synthesiser. All experiments in the present work are done in Spanish, however the process can be applied to many other languages, if an automatic phonetic transcription system is available. Results can vary from language to language though. We perform an automatic transcription of the text. Using this transcription we train demiphone-based HMM on the database and after that we use these models to perform a forced alignment over the whole database. The use of demiphone models allows to obtain phone as well as diphone boundaries. Our training system uses Baum-Welch and needs a unique transcription, so the training is done in two steps. In this first alignment we allow optional silences between words in order to find silences between words (see Figure 1). This silence model can be trained at edges of sound files or at places where they are likely to appear: punctuation marks for example. This is necessary because inter-word silences are not contained in the phonetic transcription and they would mislead the alignment.

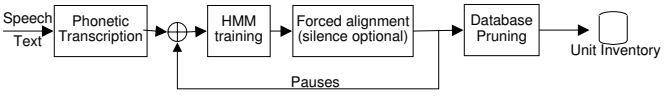

Fig. 1. Overall view of the segmentation process.

After this first step we know all the silences that may have appeared between words. Now we can train models again, and perform

This work has been partially sponsored by the European Union under grant FP6-506738 (TC-STAR project, http://www.tc-star.org) and the Spanish Government under grant TIC2002-04447-C02 (ALIADO project, http://gps-tsc.upc.es/veu/aliado)

a strict forced alignment with silence symbols where they were detected in previous alignment. The overall process is shown in Figure 1. It is possible to iterate the training and silence detection alignment processes in order to detect silences more accurately, since at every step the models will be more accurately trained.

Then, we have boundaries for every unit in the initial transcription. If some of the units were missing or some phones have been removed by the speaker spontaneously, then boundaries will be misplaced and when these units would be used in the synthesis process the output sound would contain discontinuities. Therefore, at this point we will try to detect undesired units. We will consider undesired units the ones where phonetic transcription is not correct, reading mistakes, noises and also mismatches between canonical transcription and pronunciation. In next Section 3 confidence measures used to prune the database will be deeply discussed.

3. CONFIDENCE MEASURE-BASED PRUNING

In order to detect errors, likelihoods obtained for each frame during the second alignment are used. We also do speech decoding using for each segment phones that are not the one transcribed. The ratio among likelihoods will tell us which phone has not been pronounced as transcribed.

Which units have to be kept and which ones will be removed from the inventory will have to be decided by means of a threshold. Once a unit is decided to be removed then adjacent units are also removed due to boundary effects. If a unit is incorrectly transcribed then it is not possible to accurately detect its boundaries, and adjacent units are affected too.

A variety of works have been published on using confidence measures to improve speech recognition [10] or to help other modules on natural language applications, for example, on dialogue systems. We will now use this approach to detect when some speech segments do not correspond with what has been transcribed. The sum of log-likelihoods across all frames in a unit is calculated. Log-likelihood is calculated given the transcribed model and given the best of an alternative set of models. By this method we are normalising the likelihood of the transcribed model. If one of the models in the alternative phone-set can better describe the unit, then the ratio will grow up.

We calculate a ratio between both likelihoods and we call it Transcription Confidence Ratio (TCR). It can be expressed as follows:

$$TCR = \frac{\sum_{i=1}^{N} log(p(f_i|m_t))}{\sum_{i=1}^{N} log(p(f_i|m_b))} \quad (1)$$

where f_i is the ith frame of the unit, m_t is the model corresponding to the phonetic transcription and m_b corresponds to the best aligned model from the alternative set.

This method can be seen as a classification problem where we want to classify correctly and incorrectly transcribed units. This is done as follows:

$$c(\mathbf{x}) = \begin{cases} incorrectly\ transcribed & TCR \geq th_{TCR} \\ correctly\ transcribed & TCR < th_{TCR} \end{cases} \quad (2)$$

where c is the classification function and th_{TCR} is the threshold.

Therefore, the choice of model b must be done in order to improve the separability of both classes. We present here two methods to calculate this ratio depending on how to choose model b. In both methods mentioned here m_b is the model with higher likelihood from the alternative set of models. The difference between both methods is in the set of models among which to choose.

First method, which uses the complete set of phones as alternative phone-set is presented and afterwards a method based on a restricted phone-set. The alternative phone of first method will show a high likelihood when finding a completely different phone. This is because all possible phones are available in the alternative phone-set. However, in the second approach only similar phones are taken into account, thus the alternative phone will show higher likelihood when similar phones appear. Therefore, first method is tuned to detect deletions and second method, based on a restricted set, is tuned to detect substitutions. Therefore, they could be combined. This is verified in Section 5 and also a combination is proposed and tested.

3.1. Phoneme network-based TCR

This first approach consists on using the whole phone-set as alternative phone-set. A phone network where transitions between any phone to any other are allowed is used to perform the alignment.

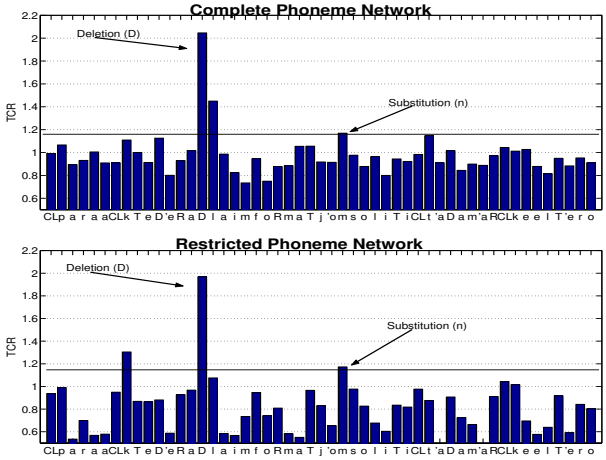

Fig. 2. Evolution of the Transcription Confidence Ratio through a sentence that contains one substitution and a deletion. TCR is shown for every phone. Transcription correspond to sentence: *"Para acceder* **D** *a la informació* **n** *solicitada teclee zero."*

Since the whole phone-set is used, the transcribed phone is contained in it. Thus, when transcription is correct we must get similar likelihood as with the forced alignment process, i.e. TCR$\sim$=1, and if transcription is not then TCR will be larger than one because the recognition process would have found a better alignment than the one given by the transcription. However, due to recognition accuracy is lower than 100% in practice TCR can be lower than one.

In Figure 2 it can be seen the variation of both models and of TCR. It can be observed how TCR goes clearly above one in the case of the deletion. In the case of a substitution then TCR is not clearly above all other phone values. An horizontal line represents a possible threshold. For some correctly transcribed units TCR goes above such threshold. This illustrates how some correctly transcribed units will be classified as incorrectly transcribed. Therefore, a trade off must be found by adjusting the threshold value.

3.2. Restricted phoneme network-based TCR

In this second approach we are using again the same framework but now a shorter list of phones is allowed. Only phones that are close to the automatically transcribed one are allowed since usually substitutions in speech concern close related phones.

In contrast with previous method, the restricted phoneme network does not contain the transcribed phone in the alternative set of phones. Thus, when transcription is correct the TCR value goes clearly below one since the best alignment of the alternative set will have lower likelihood than the transcribed one. Values for the restricted network-based method go further below one than for the first method, thus variance is larger (see Figure 2).

Then, in case of substitutions the best unit in the alternative phone-set will have lower log-likelihood than the transcribed model, thus TCR will be larger. Moreover, for deletions TCR values will be close to the ones for correct units. This happens because in such situations transcribed models do not correspond to the pronounced phone, but the unit with higher likelihood of the alternative set will not correspond neither, since the phone-set is restricted in this case.

4. EVALUATION METHODOLOGY

Evaluation is performed in order to find out whether the system proposed is able to detect units that have been either substituted or deleted by the speaker itself. Then, an evaluation framework must be established in order to have a correct objective evaluation of the system.

The method used here is based on the work done by [11]. It is not possible to introduce controlled errors in the recorded voice. Furthermore, real errors request for a big effort on classifying them. Therefore, in order to measure if the system does solve them or not, controlled errors are introduced in the phonetic transcription. This allows us to control the amount and type of errors and therefore better interpret the results.

The proposed system in both of its configurations, i.e. restricted or not, can only detect substitutions and deletions, then they have to be created by modifying the transcription. Some phones are thus substituted and some new ones are added. Note that an addition in the transcription simulates a phone that has not been said by the speaker (i.e. a deletion).

1.6% of units have been modified in the transcription in order to generate mentioned errors. Deletions have been generated by randomly inserting any phoneme in the phone-set between any two phonemes in the transcription. Substitutions have been generated by randomly substituting a phone by another one from the phone-set. Restricted phone-set for each phoneme contained all phonemes that differ from the trancribed one in one the following attributes: *articulation manner*, *articulation point*, *voice* or *vowel/consonant*.

5. EXPERIMENTS

In this section the experiments performed in order to evaluate the system proposed are detailed. Results are shown and compare across the different approaches. Specificities from both systems will lead us to propose a combination of them that improves results from both previous systems. But, first the corpus used for these evaluations is described.

5.1. Corpus

We have used a Spanish corpus recorded by a woman. The corpus is about three hours of speech. It contains short sentences, long sentences, numbers, and special phrases such as indications, locations, etc. This corpus has been semi-automatically pre-processed. It was automatically transcribed and segmented and then manually supervised. Therefore, it is considered to be correctly transcribed and segmented.

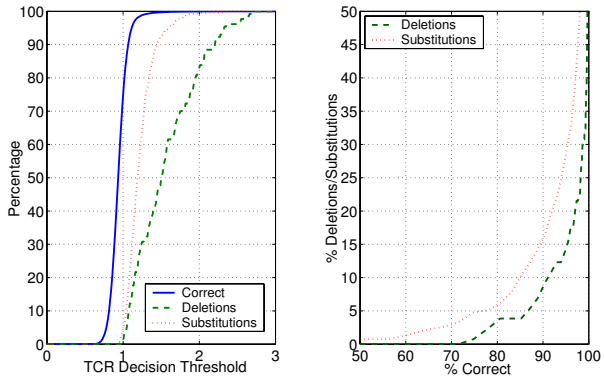

Fig. 3. Results for the system based on a **complete** phone-set. *Left*: Distribution of correct units, deletions and substitutions with respect to the TCR threshold (th_{TCR}). *Right*: Relation between percentage of correct units and the percentage of deletions and substitutions kept in the database.

5.2. Results

For the first approach, the one based on a complete phoneme network, objective evaluation has shown, as it can be seen in Figure 3, that for a threshold lower than 0.8 all deletions and substitutions are removed from the database. However, for this value only 32% of correct units would stay in the database. Since, we need a trade off between incorrectly and correctly transcribed units to be removed, if we assumed that it is reasonable to loose a 10% of a database then 84% of substitutions and 94% of deletions are removed from the database what is a significant improvement of the database.

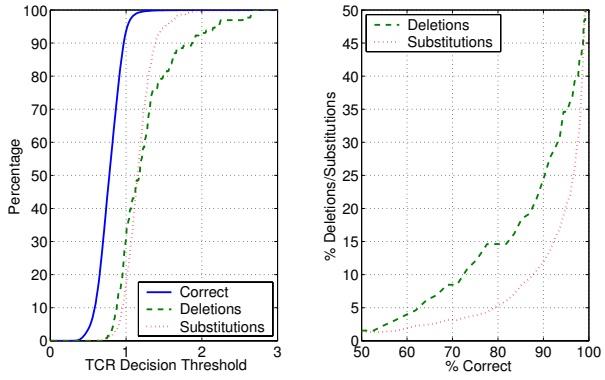

Fig. 4. Results for the system based on a **restricted** phone-set. *Left*: Distribution of correct units, deletions and substitutions with respect to the TCR threshold (th_{TCR}). *Right*: Relation between percentage of correct units and the percentage of deletions and substitutions kept in the database.

On the other hand, the approach based on a restricted phoneme network, keeping again 90% of correct units only 75% of deletions are removed while 89% of substitutions (see Figure 4). Then, it can be observed how each of both approaches better solves one of both problems (see Section 3).

Since both methods solve one of the two problems better than the other (i.e. deletions and substitutions), a combination of both methods is proposed here. A simple high level combination has been

tested. It consists on combining the TCR values coming from both methods. The resulting value corresponds to the mean of previous:

$$TCR_i^{combination} = \frac{\left(TCR_i^{complete} + TCR_i^{restricted}\right)}{2} \quad (3)$$

where $TCR_i^{combination}$ is the new TCR value for the ith phone, $TCR_i^{complete}$ is the TCR value of the ith phone corresponding to the method based on a complete alternative phone-set and $TCR_i^{restricted}$ is the TCR value of the ith phone corresponding to the restricted alternative phone-set-based method.

Results for this method combination is shown in Figure 5. Now, keeping again 90% of correct units 92% of deletions and 89% of substitutions are removed from the database. For this trade off the threshold chosen is $th_{TCR} = 1$.

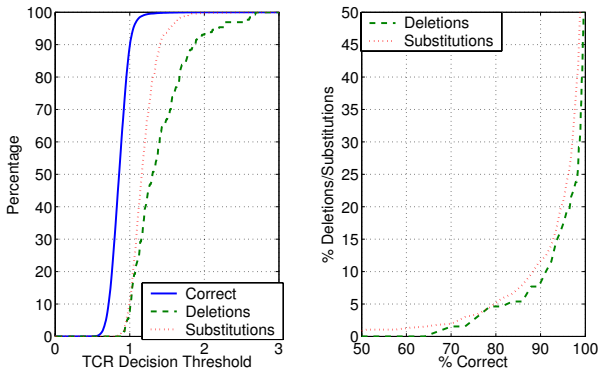

Fig. 5. Results for the **combination** of both systems. *Left*: Distribution of correct units, deletions and substitutions with respect to the TCR threshold (th_{TCR}). *Right*: Relation between percentage of correct units and the percentage of deletions and substitutions kepts in the database.

6. CONCLUSIONS

Although it is necessary a trade off between correct and incorrectly transcribed units present in the final inventory, the method proposed can clearly remove around 90% of incorrectly transcribed units by keeping at the same time 90% of the correct units present in the database.

It thus can be taken into account in the database design step. Once it is known that 10% of the inventory will be removed in order to clean the recorded database it must have been designed with 10% larger. A drawback of the presented approach is the threshold selection. It can be set by manually supervising a small part of the database, i.e. 1%. However, experimental results show that 1 is clearly more than a candidate.

Another further usage of the proposed method is in selection costs. Since TCR is related to transcription confidence it seems reasonable to use it in selection. This would make the synthesiser to unlikely use less reliable units, they would only be used where no other units were available.

The proposed method has been used in building a 10 hours voice for the first evaluation campaign within the TC-STAR[1] project. Results in such evaluation were encouraging since the voice was built in less than one month by one single person. In a Mean Opinion Score evaluation of the overall quality it reached a value of 4.1, a similar value to another commercial synthesiser that is supposed to use a more accurately built database.

[1] http://www.tc-star.org

7. REFERENCES

[1] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of ICASSP 96*, 1996, vol. 1, pp. 373–376, Atlanta, Georgia.

[2] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS system," in *Proc. of Joint Meeting of ASSA*, March 1999, Berlin, Germany.

[3] R. E. Donovan, A. Ittycheriah, M. Franz, B. Ramabhadran, and E. Eide, "Current status of the IBM trainable speech synthesis system," in *Proceedings of Eurospeech*, September 2003, Geneva, Switzerland.

[4] Murtaza Bulut, Shrikanth S. Narayanan, and Ann K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *Proceedings of ICSLP*, 2002.

[5] E. Eide, A. Aaron, R. Bakis, W. Hanza, M. Picheny, and J. Pirelli, "A corpus-based approach to <AHEM/> expressive speech synthesis," in *Proccedings of 5th ISSW*, June 2004, pp. 79–84, Pittsburgh, USA.

[6] Matthew J. Makashay, Colin W. Wightman, Ann K. Syrdal, and Aliasir Conkie, "Preceptual evaluation of automatic segmentation in Text-to-Speech synthesis," in *Proceedings of ICSLP*, October 2000, Beijin, China.

[7] Jordi Adell, Antonio Bonafonte, Jon Ander Gómez, and María José Castro, "Comparative study of Automatic Phone Segmentation methods for TTS," in *Proceedings of ICASSP*, March 2005, Philadelphia,PA,USA.

[8] John Kominek and Alan W. Black, "Impact of duration outlier removal from unit selection catalogs," in *Proceedings of the 5th ISCA Workshop on Speech Synthesis*, July 2004, pp. 155–160, Pittsburgh, Pennsylvania.

[9] Chih-Chung Kuo, Chi-Shiang Kuo, Jau-Hung Chen, and Sen-Chia Chang, "Automatic speech segmentation and verification for concatenative synthesis," in *Proceedings of Eurospeech 2003*, September 2003, pp. 305–308, Geneva, Switzerland.

[10] Eduardo Lleida and Richard C. Rose, "Likelihood ratio decoding and confidence measures for continuous speech recognition," in *Proc. of Internationa Conference on Speech and Language Processing*, October 1996, vol. 1, Philadelphia, PA, USA.

[11] Lijuan WANG, Yong ZHAO, Min CHU, Frank K. SOONG, , and Zhigang CAO, "Phonetic transcription verification with generalized posterior probability," in *Proc. of Interspeech*, September 2005, pp. 1949–1951, Lisboa, Portugal.