QUALITY CONTROL OF AUTOMATIC LABELLING USING HMM-BASED SYNTHESIS

Sathish Pammi, Marcela Charfuelan, Marc Schröder

DFKI GmbH, Language Technology Lab, Saarbrücken and Berlin, Germany
firstname.lastname@dfki.de

ABSTRACT

This paper presents a measure to verify the quality of automatically aligned phone labels. The measure is based on a similarity cost between automatically generated phonetic segments and phonetic segments generated by an HMM-based synthesiser. We investigate the effectiveness of the measure for identifying problems of three types: alignment errors, phone identity problems and noise insertion. Our experiments show that the measure is best at finding noise errors, followed by phone identity mismatches and serious misalignments.

Index Terms— Text-to-speech synthesis, unit selection, hidden markov models.

1. INTRODUCTION

Controlling the quality of automatic labelling is a topic of high interest in speech technology. In particular in unit selection and HMM-based speech synthesis systems, accurate phonetic segmentation and labelling are required to ensure quality of speech. In these systems phonetic segmentation and labelling are used for classifying and selecting appropriate units in the former and in estimating model parameters in the latter. The most precise method for labelling speech data is manual labelling by linguistic experts. This task is both time consuming and complicated, therefore automatic methods and algorithms have been developed over the last years [1, 2, 3]. Automatic labelling methods are still error-prone, so that they are often followed by a stage of manual correction, which can be directed by a confidence information that indicates which labels to verify [4].

In unit selection the quality of labelling determines the quality of units, which might be affected by a range of problems including misaligned phone boundaries, mismatches between the phones that are labelled and that are pronounced, and the presence of background noise. Estimating the quality of individual units in the database is a key issue in order to reduce the amount of manual correction effort or as a criteria to apply when choosing a unit during synthesis.

In this paper we present a quality control measure of automatic labelling. We have developed this measure on the basis of a similarity cost between two sets of phonetic segments. The first set is obtained from real speech by automatic labelling; the second set is generated from text prompts by an HMM-based speech synthesiser. One of the methods employed in automatic labelling uses a speech synthesiser to create a synthetic reference signal on which real speech can be aligned [4, 1]. In a similar way we also align a synthetic speech reference with a real speech signal, but in our case, the synthetic speech is HMM-based and the objective is not labelling but calculating a quality measure.

Our assumption is that the HMM-based generated segments will be a good approximation of the context-dependent average segment acoustics, because in HMM-based synthesis the conventional parameter generation algorithm maximises the likelihood of a given HMM making the generated trajectory close to a mean vector sequence [5]. As a result, the similarity cost is expected to spot segments that are untypical and therefore potentially problematic.

The paper is organised as follows. In Sections 2 and 3 the basic idea behind the proposed method and its technical realisation are presented. Experiments, analysis of results as well as discussion are presented in Sections 4 and 5. Conclusions and future work are presented in the final section.

2. METHOD

The basic idea behind this method is to develop a cost measure by comparing HMM based synthesis and recorded speech with their corresponding unit segment labels. Dynamic time warping (DTW) is used as a spectral comparison technique to compute a similarity match score. In particular, the match score between HMM-based (statistical model) synthesis and real speech is similar to a likelihood measure of Gaussian Mixture Models (GMMs) used in HMM-based synthesis. So, we call the match score or measure a 'statistical model Cost' (sCost).

As shown in Figure 1 the sCost is computed in several steps. As a first step, an automatic labeller estimates automatic segment labels based on recorded speech and phonetic

The research leading to these results has received funding from the DFG project PAVOQUE and from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 211486 (SEMAINE).

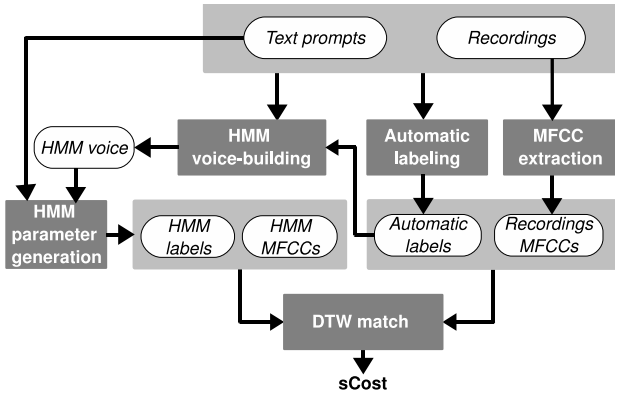

Fig. 1. Flowchart for sCost computation

transcription from text prompts. Secondly, an HMM voice is created by the HMM voice-building module using the automatic labels, generated in the previous step, and recorded waveforms. In the next steps, the HMM parameter generation module generates mel-fequency cepstral coefficients (MFCCs) and HMM predicted segment labels from the text prompts. Having similar conditions for MFCC dimension, frame size and frame-shift, MFCCs are extracted from the recorded waveforms. Finally, DTW computes an sCost by matching the MFCC feature vector sequence of the recorded speech and the MFCCs generated by the HMM parameter generation module. When aligning the two MFCC sequences their corresponding unit segment labels are taken into account.

3. TECHNICAL REALISATION IN THE MARY ENVIRONMENT

3.1. Automatic labelling

We use the automatic labeller available in the MARY voice building tools [6]. This labeller is based on the EHMM tool [3], which is well tuned to automatic labelling for building synthetic voices. Using this tool, continuous models with one Gaussian per state, left-to-right models with no skip state and context-independent models trained with 13 MFCCs are used to get force-aligned labels.

3.2. HMM-based voice building

For creating the HMM-based voices we use the programs and a modified version of the training scripts provided by HTS [7]. The modified scripts are also available in the MARY voice import tools for creating HMM-based voices for the MARY system. We use the standard training procedure (default configuration) where the spectrum is modelled by MFCC coefficients, the excitation part is modelled by log fundamental frequency (log F0) and state durations of each HMM are modelled by a multivariate Gaussian distribution.

For generating the synthetic phonetic segments and MFCC coefficients we use the MARY HMM-based synthesiser, which is a Java ported version of the hts_engine API version 0.9. During HMM-based synthesis the text analyser of the MARY system converts the text prompts into a context-based label sequence. This sequence is converted into a sequence of context dependent HMMs. State durations of each model in this sequence are estimated from the Gaussian distributions and a phoneme label sequence is generated and saved in a file. The next step is generation of parameters based on the sequence of context dependent HMMs. MFCC coefficients and log F0 values are generated using the maximum likelihood parameter generation algorithm including global variance [5]. For this experiment the generated MFCCs are saved in a file; we are not using the final vocoder stage to generate speech.

3.3. Dynamic Time Warping

DTW is a spectral comparison technique with optimal alignment to match the acoustically most similar sections between two phonetic segments. Here, an automatically labelled phone segment in the recorded speech is matched with the corresponding segment generated by the HMMs. The criterion for finding the optimal path is the Mahalanobis distance between the recorded and generated MFCC vectors, using the variance computed per phone on the recorded waveforms. sCost is computed as the sum of the Mahalanobis distance over the optimal path, divided by the number of frames in the recorded segment and in the generated segment.

4. EXPERIMENTS AND ANALYSIS

4.1. Database

The speech database used in this experiment is a German language database generated in the framework of the project PAVOQUE [8]. From this database we have selected a section of phonetically balanced neutral reading text. The duration of the corpus is around 2.6 hours and it has 1591 sentences containing roughly 73000 phone segments.

4.2. Labels used

The labels whose quality is to be assessed have been automatically generated as described in Section 3.1. As the "gold standard" regarding the quality of these labels, specially trained student assistants created a manually corrected version of these labels. Phone identity and misalignment problems, as described in the following section, are identified by comparing the automatic labels with the gold standard labels. For clarity, we use only the automatic labels in the computation of the sCost; the sole purpose of the gold standard labels in this paper is the identification of labelling problems in the automatic labels.

4.3. Error categories

We focus on three error categories that might affect the quality of labelling: alignment errors, phone identity errors and noise errors.

Alignment problems are the most common errors, where the start or end points are mis-labelled by the automatic labeller. In our database we consider as an alignment error whenever the automatic label differs from the manually corrected one. Since shorter phonetic segments are more sensitive to mis-alignments, in our study we use a normalised relative deviation (NRD) measure that allows us to analyse short and long segments equally. The NRD measure is calculated as follows:

$$\text{NRD} = \frac{\text{Relative deviation}}{\text{Duration}} = \frac{(|\text{M}_{\text{ini}} - \text{A}_{\text{ini}}| + |\text{M}_{\text{end}} - \text{A}_{\text{end}}|)/2}{|\text{A}_{\text{end}} - \text{A}_{\text{ini}}|}$$

where M_{ini} is the start time of a manually corrected phonetic segment and A_{ini} corresponds to the start time of an automatically labelled segment. The M_{end} and A_{end} correspond to the end times. NRD represents alignment errors relative to the phone duration. For example, NRD greater than 0.1 for a particular phonetic segment means that the combined mis-alignment of start and end points is more than 10% of its duration. As a simple method for classifying the severity of mis-alignments, we distinguish two groups based on arbitrary NRD thresholds: "serious" misalignments with NRD > 0.25, and "moderate" misalignments with $0.1 <$ NRD < 0.25.

Phone identity problems, the mismatch between speaker pronunciation and phonetic transcriptions. Automatic segmentation methods depend on phonetic transcriptions generated by using a pronunciation dictionary or letter-to-sound rules. Wrong lexical entries in a pronunciation dictionary or wrong predictions by letter-to-sound rules will lead to mis-alignments. Speaker specific pronunciation or slurring are also considered in this category. We consider it a phone identity error whenever an automatically predicted label differs from the manually corrected one.

Noise problems represent background noise during the recording phase, including lip smacks and breaths. Since our database has been recorded in a clean environment, we have added artificial noise to every tenth recorded utterance. We used equally white Gaussian noise and pink noise, applied in equal shares so as to maintain 5 dB and 10 dB signal to noise ratio (SNR). Table 1 presents a summary of labelling errors identified in our database.

4.4. Experiments and Results

Following the procedure described in Section 3 we have computed sCost for each phone segment in the database. The histogram of sCost over all phone segments is shown in Figure 2. If sCost is a suitable measure for identifying problematic units, the tail of this histogram should include a high proportion of problematic segments.

Type of error	No. units
Serious alignment errors (NRD > 0.25)	15335
Moderate alignment errors ($0.1 <$ NRD < 0.25)	17278
Phone identity errors	169
Noise insertion errors	7696

Table 1. Labelling errors identified in the database. The total number of phonetic segments considered in this database is approx. 73000.

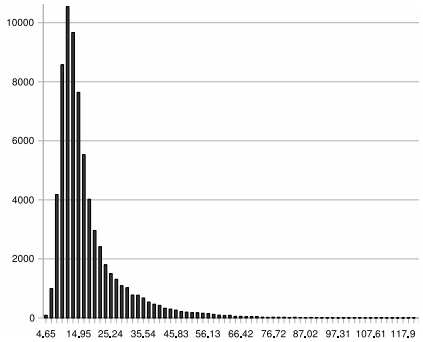

Fig. 2. Histogram of sCost over all phone segments

To test this hypothesis, we have sorted the phone segments, in descending order, according to their sCost. Table 2 presents the percentage of errors included in the first 5%, 10% and 25% of the segments sorted in this way, respectively, for noise insertion problems, phone identity problems and serious as well as moderate alignment problems. For example, it can be seen that the first 5% of the segments include 33% of all noise insertion problems, but only 3% of moderate alignment errors.

top n% of segments ranked by sCost	5%	10%	25%
Noise insertion problems	33%	59%	90%
Phone identity problems	23%	31%	59%
Serious alignment errors	10%	19%	43%
Moderate alignment errors	3%	7%	21%

Table 2. Recall statistics of the Experiment

Noise problems are clearly well spotted by sCost whereas phone identity problems and serious alignment errors are spotted moderately. sCost seems to fail to spot moderate alignment errors.

As a generalisation of Table 2, Figure 3 shows the evolution of recall when looking at the first n% of the phone segments ranked by sCost. The diagonal reference line represents the null hypothesis: an equal distribution of errors across the sorted list. Lines above the diagonal indicate that a higher-than-linear proportion of errors are found. We can observe that the noise error detection dominates the ranking: by the first 30% of the ranking most of the noise errors have been detected. Phone identity and severe alignment problems are also

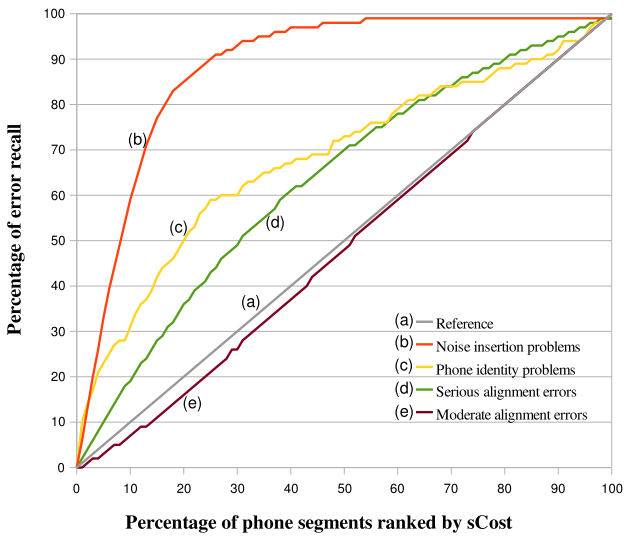

Fig. 3. Cumulative error recall as a function of phone segments ranked by sCost.

found with better-than-linear performance. Moderate alignment errors, however, are not well identified.

5. DISCUSSION

One possible reason for the failure to detect moderate alignment errors might be that our gold standard itself is not reliable enough, so that moderate deviations between gold standard and automatic labels are not reliably meaningful.

In order to understand better the performance of sCost for alignment errors, we investigated its performance for different phone classes. A small but systematic difference was found between vowels and consonants, but only for serious alignment errors: for example, the first 5% of phones ranked by sCost contain 13% of all serious vowel misalignments, but only 9% of the serious consonant misalignments. Given the importance of the Mahalanobis distance for the computation of sCost, which normalises distances by standard deviation, we compared standard deviations of MFCC coefficients between vowels and consonants. Indeed, standard deviations tend to be smaller for vowels than for consonants; for example, in our data, the average standard deviation of the second MFCC coefficient is 0.43 for vowels and 0.66 for consonants. The pattern is similar for the other coefficients. This seems to indicate a higher intrinsic variability of consonants compared to vowels. As the basic idea behind sCost computation is to compare observations with average speech acoustics, it seems natural that sCost should provide lower performance where speech acoustics are more variable.

6. CONCLUSIONS

In this paper we have described a method to estimate the quality of labelling using a statistical model cost measure, sCost, comparing recorded phones to "average" acoustics as generated by an HMM synthesis model trained on the same data. The experiments have shown that the sCost measure is effective to spot noise insertion problems, phone identity problems and serious alignment errors, but not moderate alignment errors.

One possible use of the measure could be an increased effectiveness of manual corrections if limited resources are available. When a human inspects labels in the order given by sCost ranking, more errors can be found in a given time than with simple linear inspection. Another potential use of the measure is as a quality bias in unit selection, penalising units with a high sCost over units with a low sCost.

Our future work, regarding sCost calculation, will include generation of parameters without global variance to verify if less variability on the acoustic parameters can give us a better sCost measure. Also we intend to compare our distortion measure (match score) with a likelihood directly obtained from HMMs.

7. REFERENCES

[1] F. Malfrère, O. Deroo, T. Dutoit, and C. Ris, "Phonetic alignment: speech synthesis-based vs. viterbi-based," *Speech Communication*, vol. 4, pp. 503–515, June 2003.

[2] J. Kuo, H. Lo, and H. Wang, "Improved HMM/SVM methods for automatic phoneme segmentation," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[3] K. Prahallad, A. Black, and R. Mosur, "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis," in *Proc. ICASSP 2006*, Toulouse, France, 2006.

[4] J. Kominek, C. Bennett, and A. W. Black, "Evaluating and correcting phoneme segmentation for unit selection synthesis," in *Proc. Eurospeech 2003*, Geneva, Switzerland, 2003.

[5] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[6] M. Schröder, M. Charfuelan, S. Pammi, and O. Türk, "The MARY TTS entry in the Blizzard Challenge 2008," in *Proc. Blizzard Challenge 2008*, Brisbane, Australia, 2008.

[7] K. Tokuda, H. Zen, J. Yamagishi, A. W. Black, T. Masuko, S. Sako, T. Toda, T. Nose, K. Oura, and others., "HMM-based speech synthesis system (HTS): Speaker dependent training demo version hts-2.0.1," http://hts.sp.nitech.ac.jp/, 2008.

[8] "PAVOQUE: Parametrisation of prosody and voice quality for concatenative speech synthesis in view of emotion expression," http://mary.dfki.de/pavoque, 2008.