Log Likelihood Ratio Based Annotation Verification of a Norwegian Speech Synthesis Database

Ingunn Amdal, Magne Hallstein Johnsen, and Torbjørn Svendsen

Department of Electronics and Telecommunications
Norwegian University of Science and Technology, N-7491 Trondheim, Norway
E-mail: {ingunn.amdal,mhj,torbjorn}@iet.ntnu.no
URL: http://www.iet.ntnu.no/projects/fonema/

ABSTRACT

Accurate labeling and segmentation of the unit inventory database is of vital importance to the quality of unit selection text-to-speech synthesis. Misalignments and mismatch between the predicted and pronounced unit sequences require manual correction to achieve natural sounding synthesis. In this paper we have used a log likelihood ratio based utterance verification to automatically detect annotation errors in a Norwegian two-speaker synthesis database. Each sentence is assigned a confidence score and those falling below a threshold can be discarded or manually inspected and corrected. Using equal reject number as a criterion the transcription sentence error rate was reduced from 9.8% to 2.7%. Insertions are the largest error category, and 95.6% of these were detected. A closer inspection of false rejections was performed to assess (and improve) the phoneme prediction system.

1. INTRODUCTION

Concatenation of natural speech segments is the state-of-the-art method for text-to-speech synthesis (TTS) systems. The most natural sounding systems are based on unit selection speech synthesis. This method relies on searching an annotated database of pre-recorded speech for the unit sequence which best matches a set of desired features, predicted by the TTS front-end. High quality unit selection synthesis requires that the database is annotated with accurate information about identity and position of the units. Traditionally this involves much manual work, either by hand labeling the entire database or by correcting automatic annotations. We want to make the process as automatic as possible but still achieve good quality. Automatic annotation followed by a procedure for identifying misaligned sentences, may reduce the amount of manual work.

Utterance verification is a method for assessing the output of an automatic speech recognition (ASR) system. In a TTS system automatic segmentation is usually performed using an ASR system in forced alignment mode on a predicted phone sequence obtained from running the database manuscript through the TTS front-end. The ASR acoustic models used in segmentation can also be used to compute confidence scores for the output of the forced alignment.

Utterance verification can improve the quality of unit selection databases by detecting instances where the predicted pronunciation does not match what is spoken (labeling errors), or instances where labels are misaligned. Labeling errors may be caused by the TTS front-end processing (e.g. lexicon errors and wrongly disambiguated homographs), errors in the manuscript, or reading errors (including unexpected pronunciations). Bad alignment may be caused by high speaking rate, hesitations, speaker noise, or badly trained models. Utterance verification is thus a language independent method using the acoustical characteristics of speech to correct partly language specific errors.

Using the algorithm to identify dubious sentences, these can either be discarded, or manually inspected and corrected. An advantage of a subsequent manual inspection is that we may be able to remove sources of error e.g. in the front-end. In this paper we have evaluated the log likelihood ratio based utterance verification for TTS database development presented in [1]. The data used in the experiments is a sub-set of a Norwegian two-speaker synthesis database. The sub-set is manually verified on an orthographic level to produce the "true" annotation. The method is thus tested on the transcription errors found in real data.

Section 2 gives a short overview of TTS database annotation issues including earlier work on automatic methods. The theory of utterance verification is presented in section 3. The experimental setup is explained and the results are given in sections 4 and 5. Finally, sections 6 and 7 present conclusions and suggestions for further work.

2. AUTOMATIC ANNOTATION OF UNIT SELECTION SYNTHESIS DATABASES

A TTS database usually consists of high quality recordings of a speaker (often professional) reading a manuscript designed to give the desired coverage.

Predicted phoneme sequences for the database are obtained running the manuscript through the TTS front-end. The quality of the prediction depends on the lexicon, the parser, and inevitable ad hoc rules. New words and expressions make it impossible to predict all events, and there will be ambiguities where semantic and pragmatic knowledge is needed. Words like numerals and acronyms are difficult to

To appear at Norsig-2006: www.ieee.is/norsig2006/

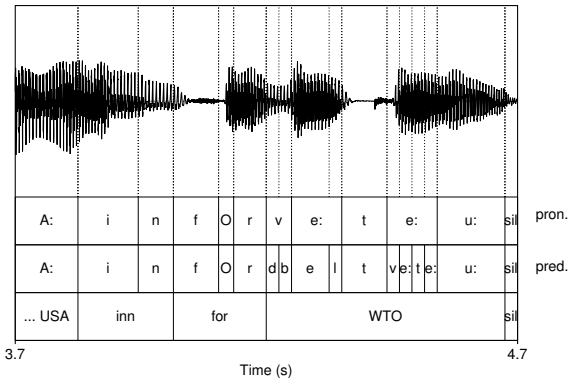

Fig. 1. *Mismatch in acronym pronunciation (top tier) and prediction (middle tier) for the phrase "(US)A inn for WTO". In Norwegian the spelling of "W" is often pronounced /ve:/ (as the letter "V") instead of /dObeltve:/.*

predict in most languages. A Norwegian example of erroneous phoneme prediction compared with the actual pronunciation is shown in Fig. 1 (transcription in SAMPA). Sloppy pronunciation, which gives rise to similar insertions in the prediction, is an important source of error.

Automatic annotation of the database is obtained using the predicted phoneme sequence and automatic segmentation procedures. These are commonly based on HMM techniques, using forced alignment of the predicted phoneme sequence. The achievable quality of the segmentation depends on the quality of the acoustic models and on the match between the predicted pronunciation and what is actually spoken. Automatic procedures make it feasible to use more data and may give higher synthesis quality than hand labeling, possibly due to improved consistency, [2].

Divergence between the predicted annotation and what is actually said is inevitable; 4% transcription errors are reported in [3]. Misaligned or wrongly labeled segments tend to have deviating acoustical characteristics. A common method is to assess segments in the initial database by computing unit statistics and remove segments far from the unit means. Another strategy is to rely on the unit selection procedure to discard these dubious segments. A third alternative is using generalized posterior probability for phonetic transcription verification as presented in [4]. This quite complex method gave an equal error rate of 8.2% on an artificially generated test set.

3. UTTERANCE VERIFICATION

Utterance verification is a well known technique used in dialog systems to assess the confidence of a speech recognition result [5]. One successful approach to utterance verification is hypothesis testing using log likelihood ratio:

$$LRT_i = \log \frac{p_i(X|H_0)}{p_i(X|H_1)} \underset{H_1}{\overset{H_0}{\gtrless}} \tau_i \qquad (1)$$

The classification of the observation X as belonging to class i, is deemed to be correct (H_0) or incorrect (H_1) depending on the value of an estimated log likelihood ratio relative to the threshold τ_i. The probabilities and the threshold must be estimated from training data. The log likelihood output of the recognizer may be used to model the correct classification. So-called "anti-models" are often used as models for the incorrect classification.

For a task on sentence level, the verification is usually phoneme based (as opposed to word based). The H_0-hypothesis consists of a sequence of phoneme identities, $\{h_0(k)\}$, and their associated time-aligned acoustic segments, see Fig. 1. Using all other phoneme models as competitors to form the anti-model, the log likelihood ratio for segment k can be computed using a smoothed average:

$$LLR_k = LL_k(h_0(k)) - \log \left[\frac{1}{N-1} \sum_{j,j\neq h_0(k)}^{N} e^{\nu \cdot LL_k(j)} \right]^{1/\nu} \qquad (2)$$

N is the number of models (phonemes). The parameter $\nu > 0$ governs the weight of the competitors. The log likelihood score for using phoneme model j on a segment X_k is $LL_k(j) = log[p(X_k|j)]$. This anti-model method has limitations because deletions in the predicted phone sequence are hard to detect. These deletions correspond to insertions in the pronunciation and should be modeled using a sequence of competitors in segment X_k.

The confidence score for a sentence is chosen as the smoothed average of the LLR scores for each segment:

$$\text{Confidence score} = \log \left[\frac{1}{L} \sum_{k=1}^{L} e^{\eta \cdot LLR_k} \right]^{1/\eta} \underset{H_1}{\overset{H_0}{\gtrless}} \tau \qquad (3)$$

L is the number of segments in the sentence. The parameter η controls the contribution from different segments. We use $\eta < 0$ as we want to emphasize the segments with low log likelihood ratio. The log likelihood scores were normalized by segment length.

4. EXPERIMENTS

The H_0-hypothesis of the utterance verification system is that the predicted phoneme sequence based on the manuscript and TTS front-end gives a correct annotation of what is actually pronounced and that the segmentation from the HMM-based system is correct. The acoustic models used to obtain the log likelihoods needed for equations (2) and (3) are the same as were used for segmentation. The HMM model topologies evaluated are thus involved in both segmentation and confidence score computation.

In a TTS database we have a more controlled task than in usual ASR settings; one speaker, known manuscript, and controlled environments. All data may be used in training since we have no need to generalize to other speakers.

4.1. The test set

The annotation verification experiments were performed on the Norwegian database FonDat1 [6]. It consists of approximately 2000 sentences, each read by two professional

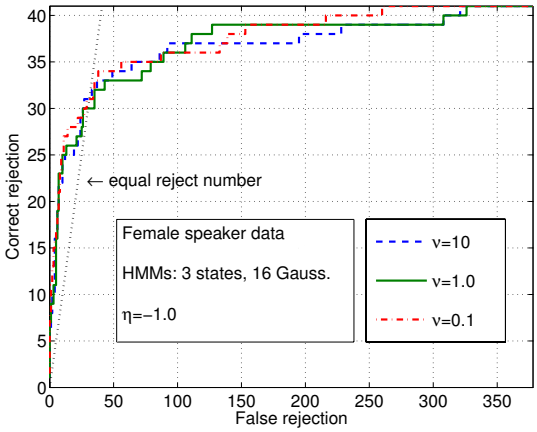

Fig. 2. ROC for test set of 419 sentences (Female speaker data: 41 sentences with annotation errors and 378 correct)

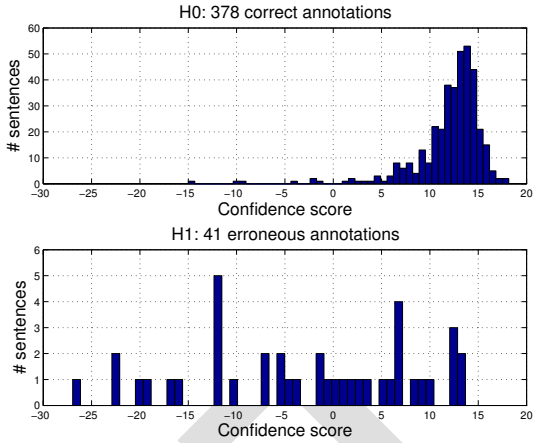

Fig. 3. Histograms showing number of sentences relative confidence score for H_0 and H_1, female speaker data

speakers, one male and one female. The test set consists of 419 sentences per speaker. All sentences contained numerals and acronyms (all capital letters). These were manually checked on an orthographic level for both speakers since the pronunciations often are difficult to predict. For the female speaker 9.8% of these sentences contained errors in phoneme prediction. The data from the male speaker (containing 11.2% annotation errors) were used to verify the robustness of the method.

4.2. The HMM system

The HTK Toolkit[1] was used to train the recognizer HMMs. The recognizer front end computed 13 static MFCCs with their first and second order derivatives every 5ms using a 15ms frame length. A speaker dependent set of 50 context independent acoustic models were produced using supervised flat start training and a phonotypical transcription created from the manuscript and a single pronunciation lexicon. All 2000 sentences read by each speaker were used as training material.

Several HMM topologies were tested. The first choice was the optimal topology for finding segment boundaries found in [7]: 7 no-skip states and observation mixture densities with 2 Gaussians. We also tested more conventional ASR topologies using 3 and 5 states and several numbers of Gaussians in the observation mixture density.

5. RESULTS

Receiver operating characteristic (ROC) curves is the traditional way of presenting utterance verification results as they show the relation between false rejections and correct rejections for different verification thresholds. For our tasks of about 10% errors it is more interesting to compare the number of rejections than the often presented rejection rates, as equal rejection rates will give 10 times more false

[1] http://htk.eng.cam.ac.uk/

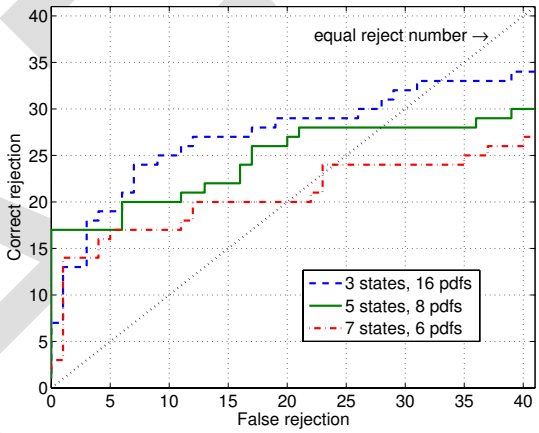

Fig. 4. ROC for chosen operating point using different HMM topologies, female speaker data

rejected sentences than correct. The operating point, i.e. the threshold for equation (3), depends on how costly false rejections are. We have a rather small database and have chosen to use "equal reject number" ERN, Fig. 2. A histogram showing H_0 and H_1 as a function of the confidence score illustrates the difficulties in defining the threshold, Fig. 3.

For all HMM topologies several experiments were performed on the female speaker data to decide the smoothing parameters in equations (2) and (3). The best values for ν were in the range 0.01–0.1 giving an anti-model close to arithmetic average of the log likelihoods of the competitors. This is the same conclusion as in [8] for utterance verification on an ASR task, but in contrast to [4], which found different settings for TTS and ASR tasks. The best values of η varies more, in most cases $\eta = -1.0$ which puts emphasis on the segments with the lowest LLR score.

5.1. Annotation error detection

Fig. 4 shows performance on female speaker data of systems with approximately the same number of parameters.

Phoneme HMM topology				
# States	# Gauss.	# Param.	ERN	EEN
7	2	1155	24	19
7	4	2261	26	19
7	6	3367	24	21
7	8	4473	24	20
5	4	1605	27	16
5	8	3185	28	17
3	8	1905	28	16
3	16	3801	33	14

Table 1. The best equal reject number (ERN) and corresponding equal error number (EEN) for different HMM topologies tested on female speaker data.

HMMs of 3 states and 16 Gaussians gave the best ERN equal to 33. Rejecting 66 sentences will then reduce the transcription errors from 41 sentences to 8 giving a reduction in sentence error rate from 9.8% to 2.7%. Table 1 gives an overview of the best ERN for the tested HMM topologies. Setting the number of false rejections equal to the number of false acceptances gives the equal error number (EEN). Too many states give worse performance, probably due to an incorrect restriction in state trajectories. More Gaussians give better performance, probably due to better modeling of feature variability. Even if this is a single speaker database in controlled environments, there are coarticulation effects that the context independent models require more Gaussians to model.

The same setting and the same confidence score threshold for the male speaker data resulted in 25 correct rejections, but 36 false rejections. The ERN for this system was 21 reducing the transcription error rate from 11.2% to 6.9%.

5.2. Error analysis

It is of great value to identify the segments within a sentence that cause a mismatch. When the log likelihood ratio from equation (2) is negative the anti-model scores better than the model. This was used in an error analysis. A closer inspection of the false acceptances confirmed the inherent inadequacy of the algorithm in detecting deletions, see Table 2. Using the ERN criterion we reject 66 sentences and thereby 95.6% of the insertions and 73.3% of the deletions. Using the EEN criterion rejecting only 28 sentences we still are able to remove 88.1% of the insertions, but only 21.7% of the deletions.

A closer inspection of the false rejections revealed some errors in the TTS front-end as well as in the manually verified transcription.

6. CONCLUSIONS

We have shown that a log likelihood ratio based utterance verification system detects real annotation errors. This is important since the quality of a unit selection based TTS system is directly related to the accuracy of the annotation.

	# Phones	Del.	Sub.	Ins.
All 419 sentences	24489	60	57	135
391 sent. accepted by EEN criterion	22616	47	27	16
353 sent. accepted by ERN criterion	20111	16	5	6

Table 2. Phoneme level transcription errors for the best HMM topology, female speaker data

The results may not only be used to discard or correct sentences, but also to identify sentences for closer inspection in order to improve the TTS front-end. The operation point for the rejection threshold depends on how many sentences we can discard. Using "equal reject number" gives a manageable number of sentences for manual analysis.

7. FURTHER WORK

The anti-model computation should be improved to handle deletions without increasing the number of false rejections. This may be done e.g. by allowing sequences of phones per segment.

A comparison with conventional methods of discarding outlier segments by computing unit statistics should be made. The practical test will be to investigate the effect on the resulting synthesis with and without the discarded sentences.

ACKNOWLEDGMENTS

This work has been supported by the project FONEMA as a part of the KUNSTI program funded by the Research Council of Norway.

REFERENCES

[1] I. Amdal and T. Svendsen, "Unit selection synthesis database development using utterance verification," in *Proc. Eurospeech 2005*, Lisboa, Portugal, 2005, pp. 2553–2556.

[2] M. J. Makashay, C. W. Wightman, A. K. Syrdal, and A. Conkie, "Perceptual evaluation of automatic segmentation in text-to-speech synthesis," in *Proc. ICSLP 2000*, Beijing, China, 2000, pp. II:431–434.

[3] X. Huang, A. Acero, J. Adcock, H.-W. Hon, J. Goldsmith, J. Liu, and M. Plumpe, "Whistler: A trainable text-to-speech system," in *Proc. ICSLP 1996*, Philadelphia (PA), USA, 1996.

[4] L. Wang, Y. Zhao, M. Chu, F. K. Soong, and Z. Cao, "Phonetic transcription verification with generalized posterior probability," in *Proc. Eurospeech 2005*, Lisboa, Portugal, 2005, pp. 1949–1952.

[5] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[6] I. Amdal and T. Svendsen, "FonDat1: A speech synthesis corpus for Norwegian," in *Proc. LREC 2006*, Genoa, Italy, 2006, in press.

[7] D. Meen, T. Svendsen, and J. E. Natvig, "Improving phone label alignment by utilizing voicing information," in *Proc. SPECOM 2005*, Patras, Greece, 2005, pp. 683–686.

[8] S. G. Pettersen, M. H. Johnsen, and T. A. Myrvoll, "Task independent speech verification using SB-MVE trained phone models," in *ITRW Robust2004*, Norwich, United Kingdom, 2004.