Unit Selection Synthesis Database Development Using Utterance Verification

Ingunn Amdal and Torbjørn Svendsen

Department of Electronics and Telecommunications
Norwegian University of Science and Technology, Trondheim, Norway
{ingunn.amdal,torbjorn}@iet.ntnu.no

Abstract

Accurate annotation of the unit inventory database is of vital importance to the quality of unit selection text-to-speech synthesis. The time consuming manual work involved in database development limits the ability to produce new voices quickly and at low cost. Automatic annotation is therefore more and more in use. Misalignments due to mismatch between the predicted and pronounced unit sequence require manual correction to achieve natural sounding synthesis. This paper proposes a new annotation assessment method using log likelihood ratio based utterance verification on the recorded database. The utterance verification is applied to detect utterances where there is a likely mismatch between the predicted pronunciation and what is actually spoken, or where an automated procedure for phonemic labelling misaligns the phone labels and the acoustic content.

In a fully automated procedure, utterances failing the verification test can be discarded. In semi-automatic procedures, the utterance verification can be applied to select utterances that need to be manually inspected, thereby reducing the manual effort. Preliminary experiments are presented that show promising figures for correct rejections.

1. Introduction

Natural sounding text-to-speech synthesis (TTS) systems are now available for many languages. The development cost of the best quality TTS is high and therefore only a limited number of voices are available. The high cost of TTS development is partly due to the amount of manual work involved. We aim for a rapid and cost efficient development of new voices. This calls for more automatic procedures in TTS development.

Concatenation of natural speech segments is the state-of-the-art method for 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. The quality is thus highly dependent on the database design [1].

Two of the main factors in the database design are the content selection and the annotation of the recorded database. Selecting the size of a unit selection database is a tradeoff between the desired variation (coverage) and the time and cost related to development, as well as search time and storage [2]. One of the first steps in database development is therefore to choose between careful design of a manuscript for a smaller database or less careful design of a manuscript for a larger database which can be pruned after recording. An related approach to the latter is to use pre-recorded databases such as audio books.

High quality unit selection synthesis require that the database is annotated with accurate information about identity and position of the units. Traditionally this involves much manual work, either by hand labelling the entire database or by correcting automatic annotations.

We want to make the process as automatic as possible but still achieve good quality. Automatic segmentation and labelling followed by a procedure for identifying misaligned sentences, may reduce the amount of manual work. Misaligned sentences can then either be discarded, or manually inspected and corrected. Manual intervention is then only needed for a subset of the recordings.

Utterance verification assesses 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. The ASR acoustic models used in segmentation can also be used to compute confidence scores for the output of the forced alignment. This paper shows how the utterance verification formalism can be used in TTS database development. This approach is new, but follows the trend of using ASR techniques in TTS systems [3].

Section 2 contains a short overview of the steps involved in unit selection database development, including annotation issues. 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 discussion, conclusions, and suggestions for further work.

2. Unit selection synthesis databases

Developing a unit selection synthesis database involves a number of steps:

1. From a large set of candidate text, remove formatting (incl. headings, tables etc.) and split the raw text into suitable chunks (sentences).
2. Expand abbreviations, mnemonics, numbers, etc. to words (text normalization).
3. Predict phonemic and prosodic content of candidate sentences.
4. Select the content of the resulting database optimizing coverage (using a greedy search).
5. Record a manuscript based database (or extract a subset of an audio book).
6. Annotate the database phonemically and prosodically.

The starting point for the annotation is an orthographic transcription whose quality depends on the text clean-up in steps 1 and 2. These steps are usually performed using ad hoc rules and may introduce (or fail to correct) errors that affect the quality of the resulting database. New words and expressions make it impossible to predict all events, and there will be ambiguities where semantic and pragmatic knowledge is needed. At every

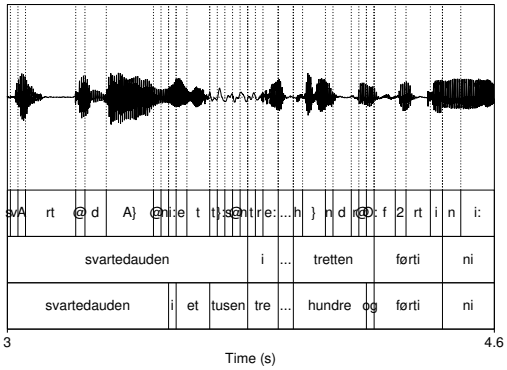

Figure 1: *Manual inspection of mismatch between number pronunciation and prediction for the phrase "Black death in 1349". Automatic word alignment in bottom tier, corrected word sequence in second tier and automatic phoneme alignment in top tier.*

step, the quality of the material is crucial. But, unfortunately, there is no such thing as error-free newspaper texts, text normalization that handles everything, perfect lexica, etc.

Numeral expressions are for example notoriously difficult to predict. An example is shown in figure 1 of a Norwegian sentence containing the phrase "...svartedauden i 1349..." ("... the Black death in 1349..."). [1] The text normalization fails to predict 1349 as a year and suggests "one thousand three hundred and forty nine" instead of "thirteen forty nine" which is spoken, causing a severe misalignment of the phone and word positions.

Careful monitoring during the recording phase can reduce the number of errors in the orthographic transcription and phonemic prediction, but the speaker may also introduce new errors when repeating utterances. We will always encounter divergence between what is predicted from the manuscript and what is actually said [4] and manual corrections are inevitable.

2.1. Annotation

The quality of a unit selection based TTS system is directly related to the accuracy of the phonemic and prosodic annotation. As shown in [5], automatic segmentation may give better synthesis quality than hand labelling, possibly due to improved consistency. Automatic procedures makes it feasible to use more data.

Automatic segmentation procedures are commonly based on HMM techniques, using forced alignment or, if multiple pronunciation alternatives are allowed, on decoding using restricted decoding networks. 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.

Misaligned or wrongly labelled segments tend to have deviating acoustical characteristics, and one strategy is to rely on the unit selection procedure to discard them. Another strategy is to assess the segments after segmentation. In [6], unit statistics were computed, and segments far from the unit means were explicitly removed from the database. An alternative is to mark dubious segments for manual inspection, as in [7], where a duration based confidence measure was used for detection.

[1] All phoneme transcriptions are given in SAMPA.

The problem of phone label errors is addressed in [8], and the solution suggested is lexicon adaptation by generating speaker-dependent lexica. Our method may be used to identify the utterances where the speaker independent lexicon fails. On the other hand improved pronunciation variation modelling will increase the accuracy of the system as the models will be better trained, reducing one source of errors.

3. Utterance verification

Utterance verification is a well known technique used in dialogue systems to assess the confidence of a speech recognition result [9]. One of the successful approaches to utterance verification is hypothesis testing using log likelihood ratio:

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

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 ration relative to the threshold τ_i. The probabilities and the threshold are not known, but must be estimated. The sum of verification errors (false acceptance and false rejections) should be minimized. 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. The anti-model may be trained specifically or we can use a combination of competitor scores. In [10] it is shown that using all competitors to form the anti-model gives the best result in the maximum-likelihood approach and almost as good performance as minimum verification trained anti-models.

For a task independent procedure, 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. 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:

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

N is the number of models (phonemes). The parameter $\gamma > 0$ governs the weight of the competitors. The log likelihood for using phoneme model j on a segment X_k is $LL_k(j) = log[p(X_k|j)]$.

The confidence score for the utterance is obtained as the smoothed average of the log likelihood scores for each segment:

$$\text{Confidence} = \log \left[\frac{1}{L} \sum_{k=1}^{L} e^{\eta \cdot LLR_k} \right]^{1/\eta} \tag{3}$$

L is the number of segments in the utterance. The parameter η controls the contribution from diffenrent segments. We use $\eta < 0$ as we want to emphasize the segments with low log likelihood ratio as they may be caused by bad alignment.

We have used the experiences from [10] for the parameter settings chosen: $\gamma = 0.1$ and $\eta = -0.1$. The log likelihood scores were normalized by segment length.

3.1. Utterance verification in database development

We wish to use utterance verification to 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; by spelling errors in the manuscript; by reading errors (or unexpected/unusual pronunciations) or by lexicon errors and wrongly disambiguated homographs. Bad alignment may be caused by deletions due to high speaking rate, hesitations, speaker or background noise, or badly trained models.

The speech database is initially passed through an HMM-based system for automatic segmentation and labelling. The labels and segment boundaries produced by this system constitute the H_0-hypothesis of the utterance verification system. The acoustic models employed by the segmentation system are used to obtain the log likelihoods that are needed for computing the utterance confidence scores using equations (2) and (3). Sentences failing the test can be discarded, or if the database is too small to be further reduced, be manually inspected and corrected.

4. Experiments

4.1. The Fonema reference database

A database for for development and assessment of Norwegian unit selection synthesis has been collected as a part of the Fonema project[2]. It consists of approximately 2000 sentences read by two professional speakers, one male and on female. The studio recordings were digitized at a sampling rate of 16kHz. The experiments reported here is using the 2000 sentences read by the female speaker.

The manuscript used in recording the database was designed using standard procedures. Newspaper texts were chosen as a starting point. Ad hoc rules were used to extract "well formed" sentences from the texts. To facilitate later phoneme prediction all sentences containing words not found in the lexicon were discarded. The remaining 75 000 senteceswere submitted to a greedy search to select approximately 2000 sentences using diphone coverage as selection criterion. The project members proof-read the sentences before the final manuscript was used for the recordings.

The instructions for reading were to use "normalized pronunciation" and a distinct way of speaking without over-articulating. The manuscript was read one sentence at the time. Each sentence was either accepted by a supervisor or the speaker was prompted to read the sentence once more.

4.2. The automatic speech segmentation system

The HTK Toolkit[3] was used to train the recognizer, perform forced alignment and to compute the log likelihood scores used in utterance verification. The recognizer front end computed 13 static MFCCs with their first and second order derivatives every 5ms using a 15ms frame length. Flat start training using the 2000 sentences read by the speaker and a phonotypical transcription created from the manuscript and the single pronunciation lexicon as training material produced a speaker dependent set of 5-state, context independent acoustic models. The observation densities were mixture densities with 8 Gaussians. This configuration was based on findings in [11].

In the initial alignment only one pronunciation per word was used. The forced alignment was performed using optional silence between the words.

[2]"http://www.tele.ntnu.no/projects/fonema/"

[3]"http://htk.eng.cam.ac.uk/"

Table 1: *Inspection of the 50 utterances with lowest confidence score.*

	type of error	instances	percentage
1	numeral	13	26%
2	punctuation	6	12%
3	POS prediction error	3	6%
4	reading error	3	6%
5	names	3	6%
6	abbreviations/acronyms	2	4%
7	segmentation error	2	4%
8	lexicon error	1	2%
9	pronunciation variation	2	4%
10	high speech rate	13	26%
11	segmentation OK	2	4%

The HMMs trained to do the initial automatic phonemic segmentation of the databases were also used for utterance verification.

5. Results

All 2000 sentences were segmented and labelled by the automatic speech segmentation system before being submitted to utterance verification. The sentences were then sorted by confidence score (eq. 3). The 50 utterances giving worst scores were manually inspected using Praat[4]. Table 1 gives a summary of the inspection results.

The first 8 rows are types of errors in alignment that we surely want to discard or correct. They sum up to 66% of the sentences. The pronunciation variation found in 4% of the sentences is also something that we want to discover, but as this experiment only uses one pronunciation per word, this problem could be reduced by using a better lexicon. The "high speech rate" sentences contain some false alarms and some sentences that require closer inspection. There are e.g. some deletions in these sentences, but we did not inspect these closer to see what segments the algorithm assigned low scores to. A closer look at the deletions due to high speech rate may be interesting in order to improve the coarticulation rules. The last row contains the undisputed false alarms.

As expected numerals are problematic, cfr. the example in figure 1. 14% of the sentences in the database contain numbers, so a strategy of inspecting all sentences containing numbers would be a tedious job. An example of acronym error is shown in figure 2: the "w" in "WTO" is read as "v" instead of "double v" which is quite normal in Norwegian for this acronym.

Examples of the pronunciation variation as well as high speech rate (e.g. sloppy speech) is shown in figure 3. The /g/-deletion in both "lørdag" (Saturday) and "ettermiddag" (afternoon) is quite normal pronunciation variation within the instruction for the actors. The deletion of a syllable (/@ t/) in "ettermiddag" is not.

6. Discussion and Conclusions

In this paper we have presented how to use utterance verification formalism to assess a unit selection synthesis database. A confidence measure is computed for each sentence that indicate whether a misalignment is present or not. The experiments show promising first results. Neither the utterance verification

[4]"http://www.praat.org/"

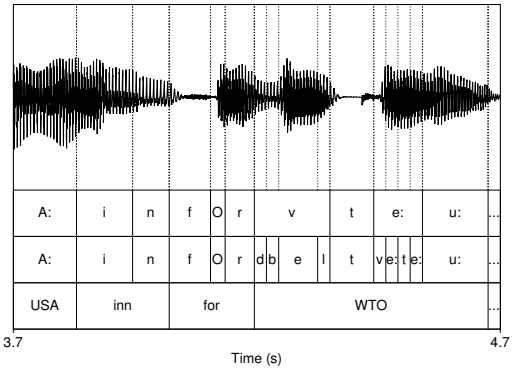

Figure 2: *Manual inspection of mismatch in acronym pronunciation and prediction for the phrase "USA in for WTO". Corrected phoneme sequence in top tier.*

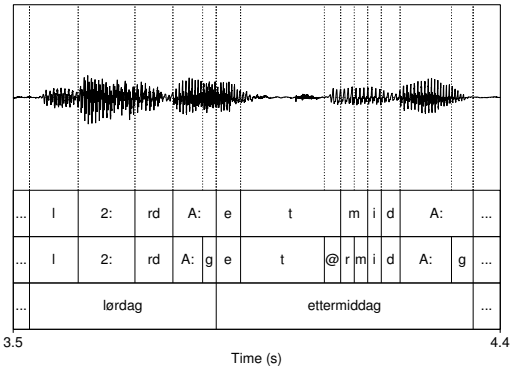

Figure 3: *Manual inspection of pronunciation variation for the phrase "Saturday afternoon". Corrected phoneme sequence in top tier.*

parameter nor the speech recognizer used in segmentation were optimized so there is most certainly room for improvement.

The settings of the utterance verification parameters were based on experiences from an ASR system, and the technique should be better tuned for the current task. The non-verbal segments may need special care, e.g. by excluding these segments in the calculation of 3. The sentences with low confidence score but no apparent segmentation error all have long silence segments.

Better performance should also be expected from using pronunciation variation in segmentation. The present algorithm may discard pronunciation variation that we want to keep. This will also give better trained models. Without pronunciation variation the acoustic models are "contaminated". Using an iterated procedure after discarding bad utterances (or correct them) should also give more precise models. Segmentation errors due to few samples of some phonemes would be reduced by using speaker independent bootstrap segmentation system.

The results may not only be used to correct/discard utterances, but also improve the lexicon, text normalization etc. It is always better to inhibit errors as early as possible in the database development steps. As for ASR utterance verification, the goal of the proposed method is to become superfluous. We want to discard as few utterances as possible.

7. Further work

First of all we would like to perform the real test of the method presented by exploring what false alarm rate we will encounter to find all (or most) correct rejections. For this we need a controlled test-set. This is also needed to find a threshold for the confidence measure.

In the Fonema project we are also working on improving the segmentation. Using the upgraded recognizer from this work should give better segmentation and more reliable utterance verification as well.

The method can be fairly easily extended to identify the problematic segments, facilitating the manual error correction further.

8. Acknowledgements

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

9. References

[1] B. Möbius, "Corpus-based speech synthesis: Methods and Challenges," University of Stuttgart, AIMS 6 (4), Tech. Rep., 2000.

[2] A. W. Black and N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Eurospeech95*, Madrid, Spain, 1995, pp. 581–584.

[3] M. Ostendorf and I. Bulyko, "The impact of speech recognition on speech synthesis," in *Proc. IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, USA, 2002.

[4] Y. Saikachi, "Building a unit selection voice for Festival," Master's thesis, University of Edinburgh, 2003.

[5] 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.

[6] 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, pp. 2387–2390.

[7] B. L. Pellom and J. H. L. Hansen, "A duration-based confidence measure for automatic segmentation of noise corrupted speech," in *Proc. ICSLP 1998*, Sydney, Australia, 1998.

[8] Y.-J. Kim, A. Syrdal, and A. Conkie, "Pronunciation lexicon adaptation for TTS voice building," in *Proc. ICSLP 2004*, Jeju island, Korea, 2004.

[9] M. G. Rahim and C.-H. Lee, "String-based minimum verification error (SB-MVE) training for speech recognition," *Computer Speech and Language*, vol. 11, pp. 147–160, 1997.

[10] 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.

[11] D. Meen, "Automatic segmentation for unit selection synthesis (in Norwegian)," Master's thesis, Norwegian University of Science and Technology, 2004.