TECHNIQUES FOR ACCURATE AUTOMATIC ANNOTATION OF SPEECH WAVEFORMS

Stephen Cox[1], *Richard Brady*[1] *and Peter Jackson*[2]

[1] School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.
[2] British Telecom Laboratories, Martlesham Heath, Ipswich IP5 7RE, U.K.

ABSTRACT

We describe techniques used in the development of an automatic annotation system for use with a concatenative text-to-speech synthesis system. The goal of the system is to generate automatically from word-level transcriptions annotations that result in synthetic speech of the same quality as that produced from hand-labelled speech. Our approach in this work has been to use the standard technique of "forced-alignment" to each utterance and to refine both acoustic and pronunciation modelling to achieve greater alignment accuracy. Acoustic models were improved by Bayesian speaker adaptation and the use of confidence measures from N-Best decodings to produce speaker dependent HMMs. Pronunciation modelling improvements involved the use of a large pronunciation dictionary containing multiple pronunciations for many words, pronunciation probabilities, the accommodation of interword silences and using information derived from existing manual annotations to guide the recogniser during decoding. At present, the system can reliably produce time-aligned phonetic alignments for UK accents in which the automatic and manual alignments agree on the segmental labelling 93% of the time. It places boundaries with an r.m.s. error of 14.5 ms from the manual boundary. Subjectively, speech produced using automatic alignments is highly intelligible if not quite as good as that produced from manual alignments.

1. INTRODUCTION

The development of high quality concatenative text-to-speech synthesis systems requires large databases of speech which are precisely phonetically annotated. In particular, "re-voicing" (the production of a new voice for the synthesiser) requires the annotation of a large amount of speech from the speaker, which is very expensive. Although automatic methods have been developed for the production of such annotations, they are generally used in the training of speech recognisers where any inaccuracies in annotation or alignment are to some extent smoothed when the data is incorporated into a statistical model. Annotation for text-to-speech synthesis demands a higher level of accuracy as any substantial errors in either labelling or alignment will noticeably degrade the quality of speech output. In this paper, we describe the development of a system which automatically annotates a speech utterance given a word level alignment of the utterance. Our approach has been to improve both the acoustic and language modelling of a standard hidden Markov model (HMM) speech recognition system (HTK V2.1) which is used in "forced alignment" mode. Although we have used objective measures of the accuracy of our system during development, our ultimate evaluation of its quality was by assessing the quality of speech it produced when compared with a system built using manually annotated speech.

2. DATA

The "Laureate" text-to-speech system [3] currently utilises manually annotated speech data taken from a database recorded at British Telecom. This consists of 239 sentences spoken by five speakers (two adult males, two adult females and one female child) and recorded to a hi-fi standard in a quiet room. The style of speech differs widely between speakers, ranging from experienced professional speakers to an untrained child speaker. A single manual transcription of each sentence was made by a professional phonetician using a 45 phone set taken from the SAMPA symbol set.

3. ASSESSMENT OF ANNOTATION ACCURACY

3.1. Figures of merit

In order to assess the quality of the annotation produced by our automatic systems, it was assumed that the manually produced alignment represented a reliable reference, and the automatic alignments were compared with this. It is well known that manual alignments show some variation from transcriber to transcriber [6] but this variation is generally small when compared with alignments currently produced by automatic systems.

After a set of automatic alignments had been made, the following figures-of-merit were estimated for the system:

1. The accuracy A of the string of phoneme symbols produced by the automatic system compared with the reference symbol string. Accuracy is defined here in the usual way as $A = 100 \times (N_C - N_I)/N$, where N is the total number of phonemes in the reference string, N_C the number of phonemes which match in the correct sequence in both the automatically generated and the reference strings and N_I the number of phoneme insertions in the automatic string when compared with the reference string. This definition is unrealistic in cases where the phoneme sequences match in the two strings but there are gross differences in the position of the segment boundaries. However, this occurs very rarely and the definition avoids the need to define arbitrary time-limits within which it may be considered that the two segmentations match.
2. Timing boundary errors were calculated for exact symbol matches only. The difference between the position of the boundaries of the reference and the automatic alignment segmentations for these symbols was estimated and the following figures of merit were derived:

Class	Symbols in Class
0	U aU, @,i, @U, eI, {, u, Q, A, E, V, O, aI, OI, I@, E@, U@, 3, r, w, j, l
1	T, v, s, f, z, h, tS, S, dZ, Z
2	k, t, D, n, d, N, m, g, p, b, M
3	#:

Table 1: The four phone classes.

(a) The mean alignment error μ_e and R.M.S. alignment error σ_e for the system.

(b) The time interval T_{90} which included 90% of the alignment errors.

The system used by Wightman and Talkin in [7] was adopted to classify and analyse alignment errors. The phoneme labels are divided into four broad classes as shown in Table 1 and the figures of merit shown above are computed for each class-class boundary. The σ_e and T_{90} values quoted here are the average figures over all transitions.

3.2. Quantisation limit on alignment accuracy

If the speech waveform is sampled at f_s Hz then clearly a segmentation boundary may be placed with a resolution of $T_s = 1/f_s$ s. However, a pattern-matching ASR system blocks the speech signal into "frames" which are computed every T_f seconds, where $T_f \gg T_s$. Hence this type of system can place a boundary with a resolution of only T_fs (typical figures for T_f and T_s are 0.01s and 0.0001s respectively). As a consequence, a system which used frame-based processing and was "perfect", in the sense that it located each boundary as accurately as possible, would suffer from a quantisation error whose R.M.S. value is given by $E_q = T_f/\sqrt{12}$ (see, for instance, [1]). We used a frame rate of 100 Hz throughout these experiments so $T = 0.01$ and $E_q = 2.9$ms. Clearly, this error can be improved by increasing the frame-rate but at a penalty of increasing processing time. In practice we found that this quantisation error was swamped by other errors.

4. DEVELOPMENT OF SYSTEMS

In this section, we discuss the development of several automatic alignment systems. Summary results for the systems are given in Table 2.

4.1. System 0: Reference system

To ascertain the best performance obtainable from an automatic aligner of this type using the data provided, we built speaker dependent HMMs for each speaker using all the available data from that speaker. The front-end representation used was 8 MFCCs with velocity and acceleration coefficients and log-energy. The HMMs were context-independent monophones with 3 emitting states and a single Gaussian distribution per state with a diagonal covariance matrix. These were used to force alignment to the reference phoneme strings. This system is unrealistic in that the models are trained on data already manually segmented and the system is supplied with the required correct phoneme string rather than a word level transcription. However, zero phoneme choice error is guaranteed and the use of the same data for training and alignment leads to excellent acoustic matching and alignment. The results were as follows: $\mu_e = -1.84$ms, $\sigma_e = 12.3$ms and $T_{90} = 24.8$ms. Figure 1 shows the mean alignment error and the T_{90} points for each class-class transition.

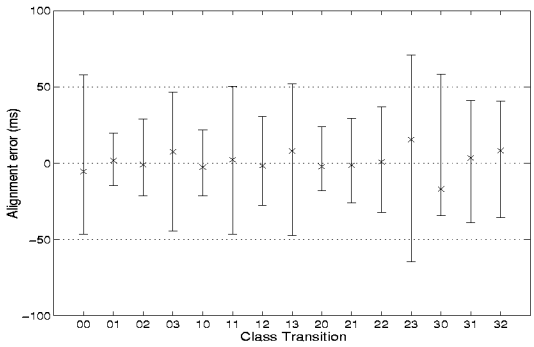

Figure 1: System 0 μ_e and T_{90} values (ms)

Figure 1 shows that there is considerable variation in the accuracy of boundary placement by the system. Fricative → voiced transitions and vice-versa (10 and 01), for instance, are quite accurately marked, whereas stop/nasal → silence transitions are subject to considerable error.

4.2. System 1: Multi-speaker models and multiple word pronunciations

A more realistic aligner was made using a single set of "multi-speaker" models built to the specification described in section 4.1 and using all the available data in the BT database. The reference phoneme string was not used for alignment purposes; rather, pronunciations for each word in the sentence were derived from the British English Example Pronunciation (BEEP) dictionary (v0.7). Many words had multiple pronunciations and these alternatives were used to build a network for the Viterbi decoder. It should be noted that this aligner is still not representative of performance expected from a real system since the data used for model training was manually-segmented and the data under alignment was also used for training. Results were as follows: $A = 94.8\%$, $\mu_e = -1.6$ms, $\sigma_e = 17.1$ms and $T_{90} = 36.1$ms.

4.3. System 2: Speaker-independent models and multiple word pronunciations

A final system must be able to produce high quality alignments for any U.K. speaker. This requires much better acoustic modelling than can be provided by the small BT database of five speakers. We therefore used the WSJCAM0 database [4] to build speaker-independent (SI) models to annotate the five BT speakers. Use of this large multiple-speaker database enabled us to build a rich set of HMMs consisting of 10 component mixture densities to model context-independent monophones. The network used was provided by the BEEP dictionary as described in section 4.2. The potential mismatch in recording conditions between training and testing was alleviated to some extent by the use of cepstral mean subtraction. Results were $A = 92.74\%$, $\mu_e = -1.58$ms, $\sigma_e = 16.89$ms and $T_{90} = 36.1$ms. This result is close to the result obtained using multi-speaker models trained upon the annotation data itself and suggests that the large number of speakers and volume of data available from WSJCAM0 was more important than the fact that it was recorded under different conditions from the BT data and was itself automatically labelled at the phoneme level.

4.4. System 3: Use of pronunciation probabilities

It was noticed that the BEEP dictionary did not contain some of the pronunciations encountered in our data and also, it gave no indication of the relative likelihood of different pronunciations. We used the manually-transcribed pronunciations in the BT database to:

1. add new pronunciations to our dictionary;
2. estimate probabilities of the pronunciations of each word.

These new pronunciations and probabilities were incorporated into our recognition network. Results were as follows: $A = 94.8\%$, $\mu_e = -1.6$ms, $\sigma_e = 17.1$ms and $T_{90} = 29.3$ms. The phoneme accuracy using these strings is extremely good, but the result must be treated with caution as some pronunciations and all probabilities were derived from material which included the material under annotation. At time of writing we are investigating deriving pronunciation probabilities independently from another database.

4.5. System 4: Modelling inter-word silences

Many utterances in the BT database contain short pauses between words. Direct use of the pronunciations provided by the BEEP dictionary does not allow for this, causing poor alignments and inappropriate silences in the resultant synthesised speech. This deficiency was rectified by allowing pronunciations to be followed by a silence. Results were $A = 92.43\%$, $\mu_e = -0.95$ms, $\sigma_e = 15.61$ms and $T_{90} = 38.2$ms. This modelling improved the synthetic speech quality by removing silences which were labelled as speech but had the negative effect of occasionally matching a silence within a stop to a silence model which can have a severe effect on the quality of the synthesized speech.

4.6. System 5: Speaker Adaptation

If some labelled data from a speaker is available, SI systems can be improved by adapting their models to the voice of a new speaker using his data. This is an attractive option for automatic alignment systems as although it may be unfeasible to manually annotate a large amount of data from a speaker, it may be possible to annotate a small amount to give a few examples of each phoneme. We experimented with using 10, 20 and 30 of the manually annotated sentences from each of the five speakers and using the Bayesian learning algorithm described in [2]. This technique reduces the mixture distribution for each state to a single Gaussian distribution whose mean and variance are adapted to the new speaker. Best results were $A = 93.06\%$, $\mu_e = 4.84$ms, $\sigma_e = 18.6$ms and $T_{90} = 37.5$ms. This represents a decrease in phoneme choice error over System 4 but the alignment error was slightly worse.

4.7. System 6: Use of reference pronunciation strings

Full modelling of the effects of co-articulation would require a complex rule-based system. However, the manually annotated reference strings supply information on pronunciation and co-articulation effects occurring in the sentences in the BT database, and this information can be used to aid alignment. The reference pronunciation strings were added to the networks generated automatically by the dictionary. During alignment, the string provided by the current speaker was removed from the network to ensure the independence of the result. This technique was used with the HMMs from System 2. Results were $A = 93.12\%$, $\mu_e = -1.29$ms, $\sigma_e = 16.26$ms and $T_{90} = 38.2$ms. As for system 3, these results pre-suppose the existence of some manually annotated speech from which these strings may be derived.

4.8. System 7: Confidence measures for bootstrapped SD models

The alignment process consists of finding the optimal path through a network of possibilities which are generated by different pronunciations. It has been shown [5] that examination of the alternative less likely paths can be used to supply a measure of confidence in which parts of the transcription represent correct decodings of the speech waveform. We configured our system to output the 100 best decodings of the sentence and compared the top decoding with the other 99. For each phoneme in the top decoding, the number of other decodings that included this phoneme in the same position was noted. The proportion of the total decodings that included this phoneme was taken as an estimate of the confidence that the phoneme was correct. A set of speaker dependent HMM's was then generated using the top decoding to provide a labelling but ignoring any data that did not have a confidence probability above a chosen threshold. These single Gaussian HMM's were then used with the pronunciation networks of System 6 to generate annotations. Results were $A = 93.42\%$, $\mu_e = -1.39$, $\sigma_e = 14.48$ms and $T_{90} = 34.4$ms. Figure 2 shows the mean alignment error and the T_{90} points for each class-class transition.

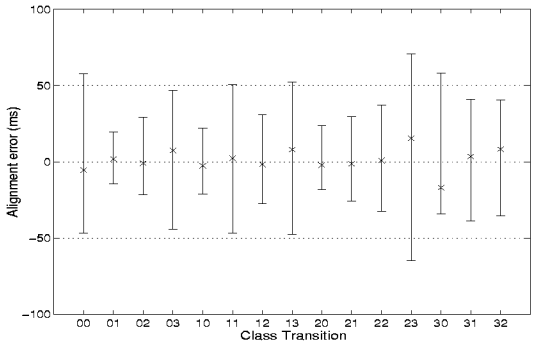

Figure 2: System 7 μ_e and T_{90} values (ms)

Comparison of Figures 1 and 2 shows that System 7 is close in performance to the reference system, System 0.

4.9. Summary of objective measurements and comments

A summary of the results for each system is given in Table 2. All time values are in ms. In Table 2, systems which are "unrealistic" to any extent (in the sense that they rely on the use of pre-existing manually annotated data) have results in italic—these are systems number 0 (the reference system), 1 and 3. Hence there is an increase in average phoneme accuracy in the performance of the successive "realistic" systems numbers 2,5,6 and 7 whilst the alignment accuracy for systems 2,4,5, and 6 is about constant. System 7 shows the best phoneme accuracy and the best alignment figures of all these systems. The system reported in [7] did not report phoneme accuracy because of the difficulty of comparing transcriptions which used different symbol sets, but it reported overall figures of $\mu_e = 2.1$ and $\sigma_e = 23.4$ ms when aligning data from the same database as used for training.

System	Phoneme accuracy		Alignment accuracy (ms)		
	Total errs	A	μ_e	σ_e	T_{90}
0	*0*	*100.0%*	*-1.84*	*12.3*	*24.8*
1	*3327*	*93.33%*	*1.29*	*15.21*	*32.1*
2	3616	92.74%	-1.58	16.89	36.1
3	*2604*	*94.77%*	*-1.60*	*17.14*	*29.3*
4	3769	92.43%	-0.95	15.61	38.2
5	3462	93.06%	4.84	18.55	37.5
6	3429	93.12%	-1.29	16.26	38.2
7	3276	93.42%	-1.39	14.48	34.4

Table 2: Summary of phoneme choice accuracy and alignment accuracy for the seven systems.

	Synthetic Voice	
Cause	Female	Male
Vowel Substitutions	15	18
Silence Insertion	4	11
Stop / Plosive	5	2
Nasal / Fricative	0	5
Other	12	12
Total	36	48
Percentage of total phonemes	0.66%	0.49%

Table 3: Pronunciation Errors in Synthesised Speech.

5. SUBJECTIVE EVALUATION

Synthetic speech based on both the automatically and manually annotated data of a male and a female speaker in the BT database was created using the "Laureate" system. Realisations of 90 sentences were generated for each of these four "voices" using System 4. Each sentence generated using automatic annotation was replayed and compared to the corresponding sentence made using manual annotation. Faulty or unusual pronunciations which occurred in the automatically annotated utterances but not in the manually annotated utterances were carefully scrutinised to determine which triphone, diphone or monophone had caused an error. The annotations were then examined and the error noted. Results are shown in Table 3.

These results indicate that phoneme substitution and insertion in the automatic annotations are the major cause of gross errors in the synthesised speech. Subjectively, the quality of the speech produced by the automatically annotated system is not quite as good as that from the manually annotated system. We suspect that some poor alignments in the automatically annotated system led to an overall poorer quality of speech than that generated by the manually annotated system, but this effect is much more subtle than the effect of phoneme substitution or insertion.

6. DISCUSSION AND FUTURE WORK

We have described some techniques to improve the accuracy of automatic annotation of speech waveforms. These included the use of a large database to build rich HMMs, speaker adaptation, the use of confidence measures to label phonemes, the use of a pronunciation dictionary containing multiple pronunciations for many words, pronunciation probabilities, the accommodation of interword silences and using information derived from existing manual annotations to guide the recogniser during decoding. Objective measurements were used to evaluate the efficacy of these techniques and a final subjective assessment revealed that synthetic speech obtained from a completely automatic annotation system was of high quality although not quite as good as that obtained from a manual system. Listening revealed that errors caused by phoneme substitutions and insertions were much more intrusive than effects from poor segmentation.

In some cases, incorrect phoneme choice was caused by the fact that the pronunciation uttered was not present in the dictionary, usually because it was pronounced rapidly and was reduced or elided with adjoining words. We therefore plan to extend this work to incorporate a model of continuous speech effects (which has been recently developed at BT) into our pronunciation networks. The use of existing manually annotated data or even of just manually derived phoneme transcriptions was found to be highly beneficial for automatic annotation accuracy and this opens up the possibility of making use of a limited amount of manually processed data for re-voicing. Speech is cheap to record but expensive to process and another area we have not yet investigated is the trade-off between using a relatively small amount of precisely (manually) annotated data and a large amount of less precisely (automatically) annotated data. Measures of confidence will be essential here to assess which sections of a waveform are reliably annotated.

7. REFERENCES

[1] S Haykin. *Digital Communications*. John Wiley and Sons, 1988.

[2] C.H. Lee, C.H. Lin, and B.H. Juang. A study on speaker adaptation of continuous density HMM parameters. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1990.

[3] J.H. Page and A.P. Breen. The Laureate text-to-speech system—architecture and applications. *BT Technology Journal*, 14(1):57–68, January 1996.

[4] T. Robinson et al. WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, pages 81–84, 1995.

[5] T. Schaaf and T Kemp. Confidence measures for spontaneous speech recognition. In *Proc. IEEE Conf. on Acoustics, Speech and Signal-processing*, April 1997.

[6] M Wesenick and A. Kipp. Estimating the quality of phonetic transcriptions and segmentation of speech signals. In *Proc. Int. Conf. on Spoken Language Processing*, pages 129–132, September 1996.

[7] C.W. Wightman and D.T. Talkin. The aligner: Text-to-speech alignment using Markovmodels. In Van Santen et al., editors, *Progress in Speech Synthesis*, pages 313–323. Springer-Verlag, 1996.