AUTOMATICALLY ASSISTED ANNOTATION OF THE AUSTRALIAN NATIONAL SPEECH DATABASE

Dong K. Kim

Speech Hearing and Language Research Centre,
Macquarie University, Sydney.

ABSTRACT - This paper describes a technique for automatically assisted segmentation and labelling of the continuous speech using CDHMM. Given the orthographic transcription of speech, the system converts this into the phonetic transcriptions which are used as the recognition network for segmentation and labelling, and then makes the forced state alignment for the speech data. Results based on the manual and the automatic transcriptions with and without the word boundary information are compared. Effects of word junction and length of state for each phone are also discussed. Our best result shows 80.08 % and 93.04 % correct boundary placement within 10 ms and 20 ms of the manual boundary respectively.

INTRODUCTION

Manual segmentation and labelling is a very time-consuming process and the inconsistency problem occurs even when carried out by skilled transcribers. The motivation for this research is to reduce the time taken to carry out phonetic segmentation and labelling. Various approaches have been proposed to automate this task.

The typical techniques use the spectrogram-reading knowledge (Zue *et al.*, 1986; Hatazaki *et al.*, 1989) and the stochastic modeling such as the Hidden Markov Models (Ljolje *et al.*, 1991; Brugnara *et al.*, 1993). The former approach provides an efficient way to represent specific knowledge associated with the speech signals, but requires extensive knowledge and each different strategy to segment and label each phone. The latter gives more flexibility for handling speech variability, and it is based on the stochastic probability of the speech events, but it does not consider various explicit knowledge representing the speech source and requires large data samples for training the reference models properly.

Currently Hidden Markov Models are largely used in automatic speech recognition and good results have been achieved (Lee *et al.*, 1991; Huang *et al.*, 1991). However, the pure phone recognition result is still not satisfactory. For the automatic segmentation and labelling purpose, it is reasonable to make assumptions, to give the correct labels, that the phonetic realisation of the given speech is already known and we only focus on locating the correct boundary of a phonetic sequence. The text-to-phonetic conversion system can be used to convert the sentence into phonetic sequence.

Based on these assumptions, Ljolje *et al.*(1991) used the trigram phonotactic model which is based on the phone sequence of the sentence to give the same transcriptions as the given input. This trigram may usually achieve this within a sentence, but it does not always guarantee to ouput exactly the same transcriptions if there is more than one phone sequence with the same pattern within a sentence. They achieved 80% boundary location within 17 ms of the manual boundary. Instead of using this trigram language model, Brugnara *et al.* (1993) built a network for each sentence to make the forced alignment of the given transcriptions with the unknown speech, and got 68.4% and 83.6% boundary placement within 10ms and 20ms respectively using the automatic procedures.

There are some problems with the automatic segmentation and labelling in this way. That is, i) text-to-phoneme(or phonetic) conversion itself may not be accurate, ii) handling silence within an utterance, and junction within a word and between words, iii) dealing with multiple alternative or optional pronunciations. In the subsequent sections these problems are discussed and the automatic procedures are investigated.

SPEECH DATABASE

The speech data were digitised at 20KHz, and segmented and labelled manually by trained the phoneticians (Croot *et al.*, 1992). 1000 sentences spoken by 4 male speakers and 1 female were used

to train the continuous density HMMs and 50 sentences for 1 male speaker, which were not included in the training data, were used as testing materials for automatic segmentation and labelling. The acoustic parameters were extracted every 5 ms using a 12 ms Hamming window and pre-emphasised. If the time interval of extracting parameters is longer than 10 ms, the accuracy will be greatly affected, and if it is shorter than 5 ms, too much calculation needs to be done. The 5 ms is a compromise of these. The parameters consist of 12 mel-frequency cepstral coefficients and 12 delta coefficients and two normalised log-energy coefficients in which each one was appended to these two streams. The speech data were trained using the Baum-Welch algorithm (Baum,1972).

The total number of phone models used are 129 which includes the basic units specified on the segmentation criteria (Croot *et al.*, 1992) and each separate model of release of oral stops, glottalised sound between vowels, and consonant-consonant sequences. Experiments were performed using the above context-independent phone models based on the continuous density HMMs with the optimal number of state and mixture for each phone model.

BUILDING NETWORKS

In this automatic procedure, the recognition network plays a role of realising the predefined phone sequence for the unknown speech. Broadly, the network can be built in two different ways. One is a word-based network in which each word has its own recognition network as in recognition. The advantage of this network is that the new words can be easily added, but it has the problem of dealing with junction between words. The other is a sentence-based network in which each sentence has a large single network. Labelling the junction between words can be well defined in this network, but it has the problem of reusability, i.e., each new network for different utterances must be created. Some important issues to be considered when building the network are described below.

Text-to-phonetic conversion

Our final goal is to have the system generate the phonetic transcriptions like the manual ones. Since our current TTS system was designed to convert text into phonemes, we have converted these phonemes into phonetic labels using rules again, and also inserted silence before and after the utterance, and between words.

Optional pronunciations

Some phonemes may have an optional pronunciation before or after that phoneme. For example, oral stops, nasals, liquids may or may not have an optional release after these. Nasals and approximants may have an voiceless part right before them, especially when /s,z,S,Z/ are preceded. The word-final /r/ sound is usually not pronounced in Australian English, but may be pronounced when followed by vowels. Some are inserted or elided.

Multiple pronunciations

Some phonemes or words have multiple pronunciations. As described in the segmentation criteria (Croot *et al.*,1992), oral stops could be one of complete or incomplete closures, or glottal sound. Glottalised vowel sounds may also be produced between two vowels or anywhere. The syllabic sounds may also be generated in /m,n,l,r/. Large numbers of unstressed vowels tend to be reduced to a weak schwa vowel. Some phonemes or words may have explicit multiple pronunciations. For instance, the first syllable of "explain" can be pronounced as /Iks/, /@ks/, or /Eks/. In the last two cases, to include these multiple alternatives in the networks, extensive investigation of all pronunciations from the dictionary as well as the actual speech database is required.

Consonant-Consonant(CC) sequence within a word and between words

There are often CC sequences which are difficult to label separately. In our continuous speech database, there are 23 different CC sequences which have more than 5 occurrences. They include stop-stop, nasal-nasal, or fricative-fricative sequence. An example of a network for stop-stop sequence (eg, /t t/ in "important to") occurring between words is (((t|tH) [th] [#] (t|tH) [th]) | (tt [th])). [ ] indicates optional, | alternative, # silence between two words, /t/ complete, /tH/ incomplete closure, and /th/ release.

Several different types of networks have been built up. They are : i) Manual Transcription-Based Sentence Network (MTSN) which is based on the manual transcription of each utterance. ii) Automatic Transcription-Based Sentence Network (ATSN) which is based on the text-to-phonetic conversion. iii) Edited ATSN (EATSN) which is based on the manually edited ATSN to include CC sequences, optional and multiple pronunciations. iv) Edited Automatic Transcription-Based Word Network (EATWN) which is based on the concatenated word network. Finally, v) Edited Automatic Transcription-Based Word Network with word boundary information (WBI) which is based on the EATWN and the word label files showing word boundary information. The word-based network does not include CC sequences occurring between words, while the sentence-based network does. The ATSN is a sort of raw network. To create the EATSN, we have manually edited the ATSNs for the 50 test sentences to allow all these variable pronunciations using the existing manually annotated data from 6 speakers and the dictionary. The EATWN was also created by the similar procedure.

ALGORITHM

The segmentation and labelling is based on the Token Passing Algorithm (Young *et al.*,1989) which has originally been developed for continuous speech recognition and is an integrated version of the One Pass algorithm (Vintsyuk, 1971) and the Level Building algorithm (Myers *et al.*, 1981).

The system takes as the recognition network the phonetic sequence of a sentence, makes the forced state alignment of these sequences with the unknown speech, finds the single best state sequence with the maximum likelihood, and then traces back to output the best matching phone sequence and the corresponding phonetic boundary. As the network includes alternative or optional pronunciations, we still need to find the single best phone sequence. If the manaul transcription is used as the recognition network, the input transcription itself will be the best and only sequence.

In addition, to see the effects of supplying the word boundary information, the above algorithm was slightly modified. The system repeats segmenting and labelling from the start time to the end time of each word until all words in the utterance have been processed, and then concatenates the results. The word-based network was used for this test because it is easier to handle word boundary information. This is to test the hypothesis that by dividing the long sentence-based speech data into the short word-based speech data, the segmentation accuracy may be improved. In this case the state sequence will be much shorter because it is based on each word, not a whole sentence.

The following diagram shows the flow of the automatic segmentation and labelling system.

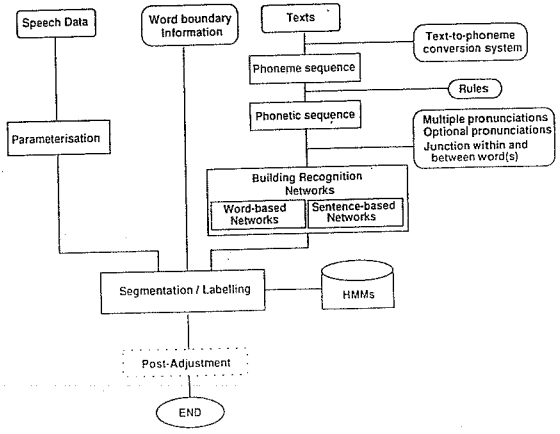

Figure 1. Diagram of the automatic segmentation and labelling system

RESULTS AND DISCUSSION

Several experiments were performed using the different kinds of recognition networks. Except for MTSN, the segmentation percentage is based on the correctly labelled phones excluding silences. The result of the automatically built network (ATSN) using text-to-phonetic conversion is roughly 2% lower than that of manually edited networks (EATSN) to cover variations in pronunciations. The difference in percentage between EATSN and EATWN indicates the effect of junction between words.

Type	Segmentation(%)				Labelling(%)				
	<= 5ms	10ms	15ms	20ms	%Corr	%Acc	%Del	%Sub	%Ins
MTSN	55.90	80.41	88.90	93.18	100.00	100.00	0	0	0
ATSN	54.09	77.92	87.00	91.60	88.44	85.47	2.97	8.60	2.97
EATSN	56.07	80.08	89.30	93.04	91.09	89.50	2.97	5.94	1.60
EATWN	55.64	79.41	88.82	92.74	89.37	86.62	3.10	7.53	2.75
WBI	66.15	83.83	91.34	94.11	88.88	86.40	3.90	7.22	2.48

Table 1. Results of segmentation and labelling

One interesting question is that if each word boundary is given to the system, how much does this contribute to improving the segmentation accuracy? It was found that this information did not improve the actual result. As an example, a particular section of the speech whose boundary was clearly placed in the wrong position was isolated and a separate file was created and then tested, but the resulting boundary was still same. The above WBI result also supports this. About 10.5 % and 4.4 % have been improved within 5 ms and 10 ms respectively, compared with EATWN. But this improvement was due to the given word boundary information which is 25.3% of the total phones (2154). This result indicates that, when all other test conditions are same, the length of speech data does not affect the result much. Rather the boundary location is more sensitively affected by other factors such as the type of HMM topology, the context-independent model, the size of HMM state, and the spectral information used.

Table 2. shows the results when using the different size of HMM state. The left-to-right 5 state HMMs for all phones degrades performance than when using HMMs with variable size to represent the differing phone durations.

Type	<= 5ms	10ms	15ms	20ms
5 state HMMs for all phones	50.13	75.27	86.26	90.83
variable state HMMs for each phones	56.07	80.08	89.30	93.04

Table 2. Results using the different size of state

Each phone has a different duration and even the same phone has large variations. In the speech recognition system, this effect may be less critical, but in the speech segmentation and labelling task, an exact modelling of duration, in terms of the size of state of each phone, is very important (Picone,1990). If the actual speech is shorter than the minimum duration (number of state multiplied by 5ms) of a phone model, and if it does not allow skip states, it will always segment at the longer position than the actual speech. Even though HMMs with skip states are modelled, it does not always improve the results. In some cases it degrades performance. The previous research (Picone,1990) calculates an optimal number of state using the clustering techniques automatically, but it does not cover all variations of duration. Our dilemma is that the size of state should not be too short or long, and the best is to get the optimal size, but it still does not solve the problem. The exact modelling of duration for the continuous speech remains unsolved yet. Since it takes a long time to train each different type of HMM, it will be helpful if an algorithm is available, which can optimise the HMM topology itself, i.e., starting from the initial prototype model and keeping on modifyng HMM topology during the training session. Most training algorithms reestimate the parameter values of the model but do not allow the number of states to be changed and the skip states to be determined during training. We have not tested this hypothesis, but

will do later. In our experiments, the optimal number of state and most suitable topology for each phone or group of phones were found through experiments.

Our final experiment shows the results based on the each different number of mixture using EATSN. The segmentation result does not fluctuate much, but the 4-mixture models reaches the best result in the labelling.

Mixture	Segmentation(%)				Labelling(%)				
	<=5ms	10ms	15ms	20ms	%Corr	%Acc	%Del	%Sub	%Ins
1	54.47	78.03	87.38	92.69	91.27	89.72	2.97	5.76	1.55
2	53.42	78.14	88.46	93.16	91.32	89.94	3.01	5.67	1.37
3	54.51	79.06	89.00	93.02	91.54	90.16	2.97	5.49	1.37
4	53.93	79.20	88.48	93.35	91.89	90.25	2.92	5.18	1.64
5	54.72	78.83	87.97	93.20	91.85	90.21	2.92	5.23	1.64
6	53.94	78.39	88.31	92.93	91.71	90.03	2.79	5.49	1.68

Table 3. Results by mixtures

Analysis of segmentation errors

Analysis of segmentation and labelling errors is based on results of EATSN whose boundary location was greater than 10ms. Stops tended to be most accurately labelled, while vowels, particularly when surrounded by sonorant sequences, were least accurately annotated. Figure 2 shows the percentage of segmentation errors (19.92%) which occurred in each different broad-class phone sequences. V-A indicates vowel-approximant or vice versa, LS errors caused by the longer size of HMM state, and ES errors of the utterance-final sound. Vowel-approximant sequences give the highest error rate (5.02%). In vowel-stop sequences, the boundary of vowels was affected by the voicing of the following stops. Approximately 1.54% was incorrectly segmented due to longer size of state of models, and other causes are coarticulation, weak sounds, and different criteria for deciding on the boundary between transcribers and machine

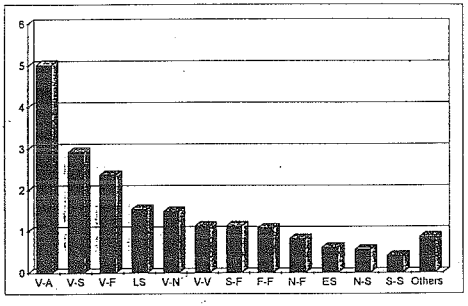

Figure 2. Percentage of segmentation errors

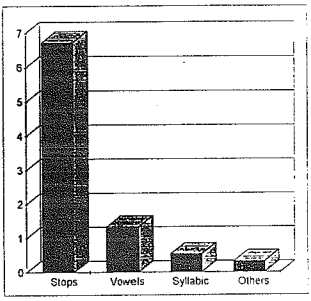

Figure 3. Percentage of labelling errors

Analysis of Labelling errors

Figure 3 shows the percentage of labelling errors (8.91%) by the broad-class phone. A major source of incorrect labelling stems from poor discrimination of complete and incomplete closure of stops. For example, the complete closure (/t/) was incorrectly labelled as the incomplete closure (/tH/) because /t/ and /tH/ have different size of state and alternative choice, but have similar acoustic feature. Many weak releases were not labelled. 6.73% was caused by this reason and 1.33 % by incorrect labelling of vowels,

and 0.53% by syllabic sounds. If we don't distinguish complete and incomplete closures, the labelling error was reduced from 8.91% to 5.89%, while the segmentation error was from 19.92% to 19.69%.

CONCLUSION

It was found that the important factors in the automatic segmentation and labelling are the size of state and its topology for each phone model. Our experiments using the CDHMMs show that the automatic segmentation and labeling tasks are feasible if we enforce this system further. To improve correct boundary location within 5 or 10 ms, more accurate modelling algorithms of duration, HMM topology, and coarticulatory effects are required. Use of context-dependant phones and post-adjustment of boundaries produced by HMMs will also improve performance.

REFERENCES

Baum,L.E.(1972) *An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov Processes*, Inequalities, 3: 1-8.

Brugnara,F.,Falavigna,D.and Omologo,M(1993) *Automatic segmentation and labeling of speech based on Hidden Markov Models*, Speech Communication,12, 357 - 370.

Croot, K., Fletcher, J., and Harrington, J.,(1992) *Labels of segmentation and labelling in the Australian National Database of spoken language*, In Proc. Fourth Australian Int. Conf. on Speech Science and Technology, Brisbane, December, 86-90.

Hatazaki,K.,Komori,Y., Kawabata,T., and Shikano,K.(1989) *Phoneme segmentation using spectrogram reading knowledge*, In Proc.IEEE Int.Conf.on Acoustics,Speech and Signal Processing, 393 - 395.

Huang,X.D., Lee,K.F., Hon,H.W., and Hwang,M.Y.,(1991) *Improved acoustic modeling with SPHINX speech recognition system*, In Proc.IEEE Int.Conf.on Acoustics,Speech,Signal and Processing, 345-348.

Lee,C.H.,Giachin,E., Rabiner,L.R., Pieraccini,R., Rosenberg,A.E.(1991) *Improved acoustic modeling for speaker independent large vocabulary continuous speech recognition*,In Proc.IEEE Int.Conf.Acoustics,Speech and Signal Processing, 161-164.

Ljolje,A., Riley,M.D.(1991) *Automatic segmentation and labeling of speech*, In Proc.IEEE Int.Conf.Acoustics,Speech and Signal Processing, 473 - 476.

Myers, D.S., Rabiner, L.R.(1981) *A level building dynamic time warping algorithm for connected word recognition*, IEEE Trans ASSP, vol 29, No. 2, 284-296.

Picone,J.(1990) *Duration in context clustering for speech recognition*, Speech Communication,9, 119-128.

Young,S.Y.,Russel,N.H.,Thornton,J.H.S.(1989),*Token Passing: a simple conceptual model for connected speech recognition systems*, CUED/F-INFENG/TR.38, Cambridge University.

Vintsyuk, T.K.(1971) *Element-wise recognition of continuous speech composed of words from a specified dictionary*, Kibernetica, vol 7, No. 2, 133-143.

Zue,Victor W., Lamel,Lori F(1986) *An expert spectrogram reader: a knowledge-based approach to speech recognition.* In Proc. IEEE Int.Conf.Acoustics,Speech and Signal Processing, 1197 - 1200.