Robust Automatic Transcription of English Speech Corpora

Ahsanul Kabir, Mircea Giurgiu
Department of Telecommunications
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
{ahsanul.kabir, mircea.giurgiu}@com.utcluj.ro

Jon Barker
Department of Computer Science
The University of Sheffield
Sheffield, United Kingdom
j.barker@dcs.shef.ac.uk

Abstract—This research assesses the ability of a Hidden Markov Model (HMM)-based method to generate an accurate and reliable automatic phone-level transcriptions for a small vocabulary speech corpus. In particular, we are interested in a system that requires only orthographic transcription of the target corpus, and can be bootstrapped from models trained on an independent phonetically transcribed corpus. The question we ask is whether reliable results can be achieved despite a large mismatch between the bootstrapping corpus (US English) and the target corpus (British English). Quality of the automatic transcriptions is judged by comparison with manual transcriptions produced by several independent transcribers. Different training strategies are compared for handling the interspeaker variability in the target corpus. The transcriptions generated from the most reliable system deviate from the average manual transcription by an average of 20 ms.

Keywords—Automatic Transcription, Hidden Markov Model, GRID Corpus, TIMIT Corpus

I. INTRODUCTION

Transcription of speech at the phone level is an essential task in speech technology research and is the first step for experiments in many areas, including language training, analysis of speech disorders, and studies of prosody and coarticulation [1, 2].

Generally, phonetic transcription is performed manually by human transcribers. There is a belief that manual transcription is more reliable and more accurate than automatic transcription. But any improvement in the quality comes with the cost of an enormous amount of time. However, time plays an important role for the transcription of large speech corpora where manual transcription requires an ample amount of effort, and in certain cases is not feasible.

When applied to spontaneous speech, phonetic labels generated by the existing state of the art automatic alignment systems differ by over 40% from those generated by phonetically trained professionals. Moreover, the boundaries generated by these automatic alignment systems differ by an average of 32 ms (40% of the mean phone duration) from the hand-labeled material [3]. This level of error dictates that manual annotation is usually necessary for detailed analysis of this type of material. However, in less variable speech domains, e.g. transcription of smaller controlled corpora often used in perceptual studies, it may be that adequate performance can be achieved using automatic techniques.

This paper looks in particular at phonetic transcription of the Grid corpus [4]. This is a small vocabulary, read speech corpus, yet it consists of a large number of individual utterances: 34,000 spoken by 34 separate speakers. Given that, even with the best transcription tools, manual transcription at the phone-level takes many times real time, hand transcription of the entire Grid corpus would be prohibitively expensive.

The paper evaluates a system based on a standard hidden Markov models (HMM) approach. In such approaches, phone-level models are first generated usually by either training on some hand transcribed subset of the corpus, or by using a ‘flat start’ where phone boundaries are initially estimated to be evenly spaced within the utterance. Automatic transcriptions are then generated using a procedure known as “forced alignment”: The phoneme sequence representing the utterance is generated from the given orthographic transcription. This is achieved by looking up the appropriate phoneme sequence for each word in a pronunciation dictionary, and then concatenating these sequences (optional silence models may be placed between words to model potential inter-word pauses in the utterance). An utterance-level HMM is then constructed by concatenating phone-level HMMs in the correct sequence. The phoneme boundaries can then be estimated by using the Viterbi algorithm to find the most likely state sequence through the HMM.

This paper examines the performance of a fully automatic transcription system based on the forced alignment approach described above. In particular we assess the performance of an automatic system that operate without access to manual phone-level transcription of any of the Grid data, but instead initialize phone models using a large hand transcribed but acoustically mismatched corpus (US English rather than British English).

II. SPEECH MATERIALS

Speech data has taken from TIMIT corpus (American Pronunciation) and GRID corpus (British Pronunciation) for the experiment. TIMIT corpus has the time alignment at the phoneme level but the GRID corpus does not. The TIMIT corpus will be used to get the initial estimates of the parameters...
of HMM; the target is to get the time alignment for the GRID corpus using forced alignment.

A. TIMIT Corpus

The TIMIT corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition (ASR) systems [5]. It contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers (492 male speakers, 192 female speakers) from 8 major dialect regions of the United States. A speaker's dialect region is the geographical area of the United States where they lived during their childhood years [5]. Each speaker read 2 dialect sentences, 5 phonetically-compact sentences and 3 phonetically diverse sentences. The dialect sentences provide dialectal variants, phonetically compact sentence mean to have a good coverage of pairs of phones and phonetic contexts, and phonetically-diverse sentences add diversity in sentence type and phonetic contexts.

B. GRID Corpus

GRID is a large multi talker audio visual sentence corpus to support joint computational-behavioral studies in speech perception and automatic speech recognition [4]. It contains a total of 34,000 sentences of high quality audio and video (facial) recordings, 1000 sentences spoken by each of 34 speakers (18 male speakers, 16 female speakers). All speak British English as their first language. All but three participants had spent most of their lives in England and together encompassed a range of English accents [4]. Two speakers grew up in Scotland and one was born in Jamaica. Grid provides a greater variety and is large enough to meet the training requirements of ASR systems.

III. SYSTEM OVERVIEW

HMM-based approaches adopted from ASR are most widely used for automatic segmentation providing a consistent and accurate phone labeling scheme [6]. There are two phases in this approach, namely HMM training, and Viterbi alignment for unit segmentation. In the first phase, each phone is defined as an HMM, and then trained with a given phonetic transcription and its corresponding feature vector sequence [6]. However, there has always been a question on how to get the initial estimates of HMM parameters. The better it initializes, the better it performs. Figure 1 shows the working method for generating automatic transcription.

A. Feature Extraction

Parameterization of the raw speech waveforms into sequences of feature vectors are performed by the use of same configuration file for both TIMIT corpus and GRID corpus. Feature vector contains 15 Mel Frequency Cepstral Coefficients (MFCC) and real energy.

B. TIMIT Training

To get the better initialization of HMMs, first we trained TIMIT speech corpus at the phone level. These trained TIMIT phone models are then adapted into our target corpus, GRID corpus.

C. GRID Adaptation

Firstly a transcription file is supplied at the word level for GRID corpus without time aligned information. Then phone transcription from words is carried out by looking up a British pronunciation dictionary, namely BEEP. Then previously trained TIMIT phone models are used to train as well as adapt the GRID corpus. GRID training is subdivided into two phases: Speaker Independent (SI) Training, and Speaker Dependent (SD) Training.

Entire GRID corpus (34000 sentences) is used to train speaker independent phone models. Speaker independent training continues until there are no significant changes in the recognition results. The phone models resulting from the previous iteration are used as the input for HMM next initialization and re-estimation.

There are 1000 sentences per speaker which are used to generate speaker dependent models. First iteration of speaker dependent training used the last models trained in speaker independent manner. Subsequent iterations use respective speaker dependent models until there are no significant changes in the recognition results.

D. Automatic Transcription

Now the system is trained properly and has sufficient knowledge about GRID corpus. Therefore, the recognition network would be able to operate in the forced alignment mode properly. The standard Viterbi algorithm is used for this purpose. We were mainly interested in three kinds of automatic transcription to determine the accuracy, reliability and usability of these transcriptions through statistical analysis, namely transcription generated a) by TIMIT phone models, b) by speaker independent GRID phone models, and c) by speaker dependent GRID phone models.

IV. EXPERIMENTAL RESULTS

The performance of the system is evaluated with manually annotated speech of GRID corpus. For this purpose, we took few randomly selected sentences which are uttered by the best speaker and the worst speaker according to the performance of the recognizer, specifically speaker of id23 and speaker of id1. Then 5 researchers (transcribers T1-T5 in Figure 3 and Figure
4) working in the field of phonetics transcribed these 10 sentences using PRAAT which is very popular software among the phoneticians for speech processing. A series of statistical analysis is carried out considering start timing and end timing of all the phonemes associated with these sentences to determine the accuracy, reliability, and usability of the generated automatic transcription. One researcher transcribed those sentences twice but in different time (approximately after one month). One of his transcriptions has been taken as the reference for all these analysis.

A. Evaluation of Acoustic Model

Recognition rate is 80.10% when testing is carried out in a speaker independent manner and with the insertion of the optional short pause among the words of sentences. But it is increased substantially when testing is carried out in a speaker dependent manner. Best performance is achieved for speaker of id23 and speaker of id9 with a recognition rate of 97.53% and 97.20% respectively where worst performance is noticed for speaker of id1 and speaker of id20 with a recognition rate of 89.06% and 90.47%. Similarly, when speech is assumed continuous (no short pause) recognition rate dropped slightly when testing is carried out in a speaker independent manner but there is almost no variation when testing is carried out in a speaker dependent manner. Figure 2 shows the recognition rate for all the phonemes when testing is carried out in speaker dependent and speaker independent manner with the insertion of optional short pause or without short pause. It necessarily signifies that TIMIT phone models are adapted successfully to GRID corpus. It also proves that American pronunciation could be a successful initial estimates of HMM parameters for British pronunciation.

B. Inter transcriber Variability

Series of statistical analysis have been carried out to determine inter transcriber variability. Arithmetic and quadratic mean (root mean square), variance, standard deviation and standard error have been calculated with respect to reference.

Figure 2. Recognition of phonemes (Speaker Independent and Speaker Dependent)

Transcription varies significantly from transcriber to transcriber and standard deviation is higher among the transcribers. It implies that manual transcription is not as stable as like the automatic transcription. Figure 3 and Figure 4 shows arithmetic and quadratic mean, variance, standard deviation and standard error among professional transcribers with respect to reference respectively.

C. Evaluation of Transcription

Statistical analysis has been carried out in a similar manner to evaluate the generated automatic transcription where arithmetic and quadratic mean, variance, standard deviation and standard error have been calculated with respect to reference.

Figure 3. Arithmetic mean, variance, standard deviation, and standard error of clear speaker (id23, black bars) and unclear speaker (id1, white bars) among the transcribers

Figure 4. Quadratic mean (root mean square), variance, standard deviation, and standard error of clear speaker (id23, black bars) and unclear speaker (id1, white bars) among the transcribers
and quadratic mean should be able to evaluate the transcription properly because only absolute value is considered for this analysis.

Speaker dependent models provide the best transcription followed by speaker independent models, manual transcription, and TIMIT models. Therefore, it could be said that significant improvement has been made after training GRID corpus. It is also noticed that human transcribers are more consistent than machine to transcribe unclear speaker. Arithmetic and quadratic, both variances tend to be very high for the human transcribers for transcribing clear speaker. In contrast, the performance of each model is pretty much consistent for clear speaker.

![Figure 5. Arithmetic mean, variance, standard deviation, and standard error of clear speaker (id23, black bars) and unclear speaker (id1, white bars) compared with models and manual transcription](image)

![Figure 6. Quadratic mean (root mean square), variance, standard deviation, and standard error of clear speaker (id23, black bars) and unclear speaker (id1, white bars) compared with models and manual transcription](image)

V. CONCLUSION

Still it is hard to say that automatic transcription is better than the manual transcription. But certainly our approach proved that automatic transcription performed well over the GRID data in compare to manual transcription considering inter transcriber variability and consistency among the transcribers. It is sure that automatic transcription is consistent and definitely it will save enormous amount of human efforts as well as time. Surely this technique will have an impact in future, especially in transcribing large corpora like GRID.

ACKNOWLEDGMENT

This work has been carried out in the frame of the EC-funded Project Marie Curie Research Training Network MRTN-CT-2006-035561 S2S (“Sound to Sense”, www.sound2sense.eu).

REFERENCES