Automatic Transcription and Speech Recognition of Romanian Corpus RO-GRID

Mircea Giurgiu and Ahsanul Kabir

***Abstract*—The results reported in this paper assess the ability of Hidden Markov Model (HMM) based method to generate accurate and reliable automatic phone-level transcriptions for a small vocabulary speech corpus such as RO-GRID. The system requires only orthographic transcription of the target corpus, and can be bootstrapped from models trained just on few amount of data in the transcribed corpus. For this purpose, an automatic time-aligned phone transcription toolbox has been developed and tested on the Romanian corpus and also validated on an English corpus. The quality of transcriptions is judged by evaluating the statistical parameters of the error between the automatic and manual transcription. The transcriptions generated from the most reliable system deviate from the average manual transcription by an average of 20 ms. The system is also able to convert the generated transcription from HTK format into PRAAT format for further manipulation of the speech signal.**

***Keywords*—Automatic speech transcription, Hidden Markov Models.**

I. Introduction

An essential task in speech technology research is the transcription of a speech corpus at the phone level in order to further experiment various tasks in practical application areas such as Automatic Speech Recognition (ASR), text to speech synthesis (TTS), language training, analysis of speech disorders, and studies of prosody and co-articulation [1, 2]. Manual transcription has a strong impression among the users that it is more dependable as well as more precise than automatic transcription. Unfortunately, manual transcription of large amount of speech data is very laborious and time consuming. Therefore, well-designed, easy to use software tools are absolutely essential to make the task of transcription of large amount of speech data less cumbersome.

For the large vocabulary systems, existing state of the art automatic transcription systems vary significantly (over forty percent) from manual transcription yielded by phonetically trained professionals. 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. In the case of RO-GRID corpus, although it contains a simple vocabulary, it still provides a greater variety and enough amount of speech signal to meet the training requirements of ASR systems.

The paper evaluates a system based on a standard hidden Markov models (HMM) approach. The phone-level models are first generated by training on only 5% of the hand transcribed subset of the corpus. Automatic transcriptions are then generated using "forced alignment" procedure: 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. An utterance-level HMM is then constructed by concatenating phone-level HMMs in the correct sequence. The phoneme boundaries are then estimated by using the Viterbi algorithm to find the most likely state sequence through the HMM.

Apart from the research challenge described above, there is a simple but useful and fair demand to import the transcribed data in PRAAT for the manual verification and manipulation of the speech signal. Inter transcriber variability tends to decrease when they are supplied a sample transcription for the purpose of verification. This research tries to answer this call, too.

The rest of the paper is organized as follows: Section II describes the features of the speech corpus used in this research, Section III presents the implemented system for the automatic transcription, Section IV gathers all relevant experimental results and Section V presents the conclusions.

II. The Speech Corpus

The RO-GRID corpus, the target of the automatic transcription in our research, is influenced in design by the English GRID corpus which has sentences of the form <command: 4> <color: 4> <preposition: 4 words> <letter: 25> <digit: 10> <adverb: 4> (eg. "bin blue at f 2 now") [4, 5]. This is because it will be used for cross language studies, too.

Eight male and three female speakers aged 20-28 years contributed to the corpus. Speakers are undergraduate students and PhD students at the Technical University of Cluj-Napoca. All, but one speaker is bilingual who has a family connection with Hungary. He was born in Romania, had spent his life in Romania but also speaks Hungarian. The mean age of the population is 25 years.

Recordings were made in a reasonably quiet TV studio under automated computer assisted control.

Manuscript received February 21, 2012. This work was supported in part by the Grant MC RTN 035561 of the EU.

M. Giurgiu is with the Telecommunications Department in Technical University of Cluj-Napoca, 26 Baritiu Str., 400027 Cluj-Napoca, Romania (phone: +40 264 427271; fax: +40 264 591689; e-mail: Mircea.Giurgiu@com.utcluj.ro).

A. Kabir was with Telecommunications Department in Technical University of Cluj-Napoca. He is now with the School of Computing and Mathematical Sciences, University of Greenwich, London SE10 9LS, UK. (e-mail: Ahsanul.Kabir@com.utcluj.ro).

TABLE I
THE STRUCTURE OF RO-GRID CORPUS

Command	*Color*	*Prep*	*Letter*	*Digit*	*Adverb*
vezi	negru	la	p t d g j	0-9	putin
(look)	(black)	(at)	b v h o		(few)
muta	verde	de	u		agale
(move)	(green)	(by)			(slowly)
pune	bronz	in			acolo
(put)	(bronze)	(in)			(there)
sari	auriu	cu			afara
(jump)	(golden)	(with)			(outside)

Sentences (see Table I for the vocabulary structure of RO-GRID corpus) were presented on a computer screen placed in front of the participants and had 5 seconds to speak every sentence. Participants were advised to speak in a natural manner as if they are used to communicate with others [6]. Recordings were divided into five recording sessions for each participant including the repeat session.

A total of 100 sentences made up each of the first four recording session and the repeat session was made up depending on the number of mistakes made by the individual participant so that were necessary to repeat. Recordings were completed by more or less 50 minutes for each participant, hence more than 9 hours of speech material is available.

III. THE AUTOMATIC TRANSCRIPTION

HMM-based approaches adopted from ASR are most widely used for automatic segmentation providing a consistent and accurate phone labeling scheme [7, 8]. There are two phases in this approach, namely HMM training, and Viterbi alignment for unit segmentation. For this purpose, the corpus has been divided into two sets where 5% of the corpus (15 sentences for each speaker, giving a total of 165 manually transcribed sentences from the total of 4400 sentences existing in the corpus) belongs to the train set and the rest (4235 not transcribed sentences) belongs to the test set. However, there has always been a question on how to get good initial estimates of HMM parameters. The better it initializes, the better it performs. Train set covers every word, syllable and phoneme of the entire corpus and has been transcribed manually in the word, syllable, and phoneme level. For the acoustic processing, the feature vectors contain 39 components: MFCC (Mel Frequency Cepstral Coefficients), the energy and their delta and acceleration values.

First, the initial HMM models have been generated by training the train set using hidden Markov model toolkit (HTK). For this purpose, a five-state prototype model (three emitting states from left to right and one Gaussian PDF per emitting state) is defined. Then prototype model is updated with global speech means and variances. Therefore, phoneme models are initialized and then re-estimated. And finally, trained phoneme models are combined into a single master macro file for embedded re-estimation in order to update all the phoneme models at once and that is the end of the initial training. Therefore, the main goal behind the initial training is to adapt these trained phoneme models into target corpus (RO-GRID corpus) which is to be transcribed.

After the initial training, the train set has been bootstrapped from these models and trained in speaker independent (SI), as well as speaker dependent manner (SD). Finally, forced alignment was performed by using SD models in order to provide transcription of the test set. Finally the automatic transcription was also converted into a file with a structure that can be imported into PRAAT for further processing.

IV. EXPERIMENTAL RESULTS

In order to evaluate the performance of the automatic transcription system, the segmentation results have been compared with those generated by manual transcription, and therefore the statistical parameters of the transcription error are calculated. For this purpose, we took 10 randomly selected sentences which are uttered by the best recognized speaker and the worst recognized speaker according to the performance of the automatic recognizer. Then 5 transcribers (T1-T5) in Figure 3 and Figure 4 working in the field of phonetics transcribed these sentences using PRAAT. A series of statistical analysis are carried out considering start timing and end timing of all the phonemes associated with these sentences to determine the accuracy, and reliability of the generated automatic transcription.

We also have evaluated the speech recognition performance with the trained models. Recognition rate is 85.99% when testing is carried out in a speaker independent manner (Fig. 1 and 2) 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 (Fig. 3). Best performance is achieved for speaker of id6 and speaker of id10 with a recognition rate of 97.60% and 96.64% respectively, where worst performance is noticed for speaker of id1 and speaker of id8 with a recognition rate of 91.65% and 92.62%.

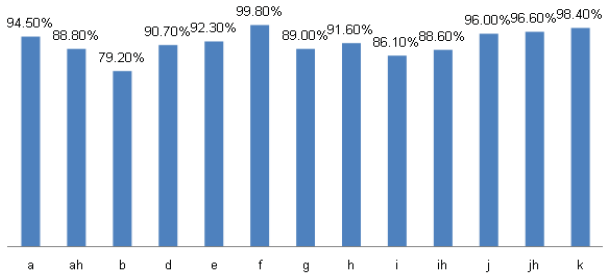

Fig.1. Speaker Independent (SI) recognition of phonemes (first part of phonemes)

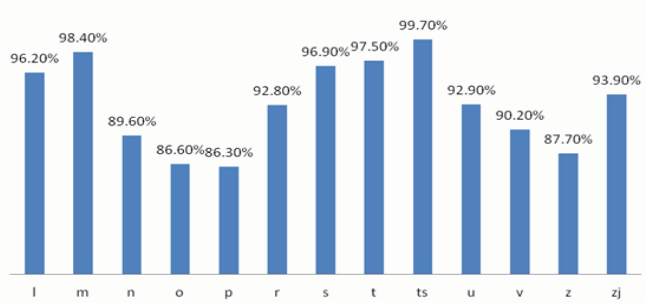

Fig.2. Speaker Independent (SI) recognition of phonemes (second part of phonemes)

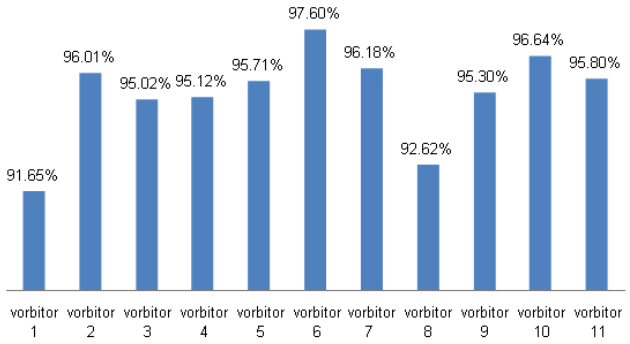

Fig.3. Speaker Dependent (SD) recognition of phonemes

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. Statistical analysis of the transcription error has been carried out to determine inter transcriber variability on the transcription of GRID corpus. Arithmetic and quadratic mean (root mean square), variance, standard deviation and standard error have been calculated with respect to reference. 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. Fig. 4 (top, down) shows arithmetic and quadratic mean, variance, standard deviation and standard error among professional transcribers with respect to reference respectively.

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.

Fig 5 (top and down) shows arithmetic and quadratic mean, variance, standard deviation and standard error of the transcription error respectively. Though arithmetic mean of the transcription error is the least for manual transcription, it has very little significance. Because positive and negative, both errors are simply added here. However, quadratic mean of the transcription error is the least for speaker dependent models 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, and manual transcription. Therefore, it could be said that significant improvement has been made after training RO-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.

Generated transcriptions are also compatible to import into PRAAT. It is possible to verify these transcriptions manually in PRAAT, modify them and save them. The toolbox is also able to get back the transcription into HTK format for further

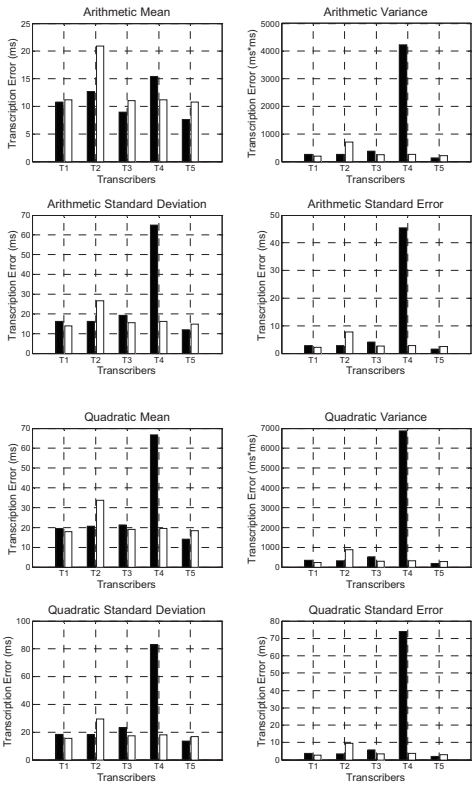

Fig. 4. Arithmetic (top) mean and quadratic (down) mean, variance, and standard deviation, of clear speaker (black bars) and unclear speaker (white bars) among the transcribers

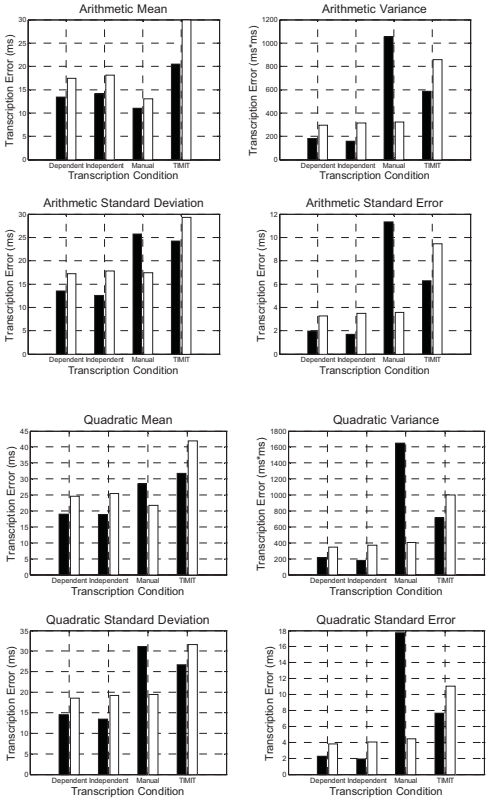

Fig. 5. Arithmetic (top) mean and quadratic (down) mean, and variance, of clear speaker (black) and unclear speaker (white) compared with models and manual transcription.

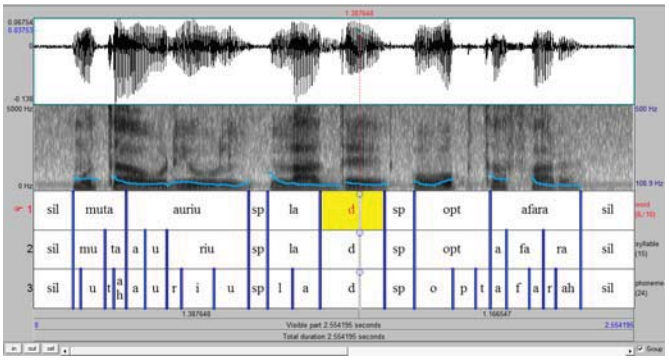

Fig. 6. Transcription for the utterance "muta auriu la d opt afara" is imported into PRAAT and can be compared to those produced manually.

use by HTK. Fig. 6 shows an automatic transcription imported in PRAAT for manual verification.

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 RO-GRID data in comparison 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.

However, it has a limitation that it requires word level transcription (without time aligned) as input and it generates time aligned phone level transcription as output. It can not be clearly said that the quality of automatic transcription is similar to manual transcription, but it is consistent and requires minimum human effort. It could be a starting point for the human transcribers as it allows importing generated automatic transcription in PRAAT for manual verification, especially while transcribing large speech corpora.

References

[1] J. P. Hosom. "Automatic phoneme alignment based on acoustic-phonetic modelling," in *Proc. ICSLP*, 2002, pp. 357-360.

[2] K. Sjolander. "An HMM-based system for automatic segmentation and alignment of speech," in *Proc. PHONUM 9*, 2003, pp. 93-96.

[3] Chang, S., Shastri, L and Greenberg, S. "Automatic phonetic transcription of spontaneous speech (American English)," in *Proceedings of the International Conference on Spoken Language*, Vol. IV, pp. 330-333, 2000.

[4] M. Cooke, J. Barker, S. Cunningham and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition." *Journal of the Acoustical Society of America*, vol. 120, 2006.

[5] J. Barker and M. Cooke, "Modeling Speaker Intelligibilty in Noise." *Speech Communications,* vol. 49, 2007, pp. 402-417.

[6] A. Kabir, and M. Giurgiu, "A Romanian Corpus for Speech Perception and Automatic Speech Recognition". *Proceedings of ISPRA 2011,* 20-22 February 2011, Cambridge, UK, pp. 323-327.

[7] Y.J. Kim and A. Conkie. "Automatic Segmentation Combining an HMM-Based Approach and Spectral Boundary Correction," in *Proc. ICSLP*, 2002, pp. 145–148.

[8] A. Stolcke and E. Shriberg. "Automatic linguistic segmentation of conversational speech," in *Proc. ICSLP*, 1996, pp. 1005-1008.