AUTOMATIC RECOGNITION OF SPONTANEOUS SPEECH DIALOGUES

Mauro Cettolo, Daniele Falavigna

ITC-Irst, Istituto per la Ricerca Scientifica e Tecnologica
via Sommarive 18
38050 Povo, Trento - ITALY
{cettolo,falavi}@itc.it

ABSTRACT

Some approaches for coping with the problem of recognition of spontaneous speech dialogues are presented. Starting from a HMM-based system, developed for dictation tasks, some modifications are introduced at the acoustic level and in the language model. Acoustic model parameters are modified to account for speaking rate variations, and specific models of extra-linguistic phenomena are defined and added to the language model. Different acoustic models are managed by a single recognizer through their integration into a multi-model search space. Experiments and evaluations were conducted on a spontaneous dialogue corpus collected at our laboratory.

1. INTRODUCTION

This paper concerns the assessment of a set of modifications applied to a HMM-based continuous speech recognizer developed at ITC-Irst, for dictation tasks, to improve its performance on a corpus of spontaneously uttered human-human dialogues. As the performance of Automatic Speech Recognition (ASR) systems is largely affected by typical phenomena of spontaneous speech, such as noises, speaking rate variations, false starts etc., it is advisable to take them into account when developing the recognizer.

Some solutions are herein proposed to increase ASR robustness with respect to the above phenomena. Explicit models are used to cover some extra linguistic phenomena (breaths, coughs, filled and silent pauses). Model parameters are properly modified to take into account speaking rate variations. This approach is applied to gender independent as well as to gender dependent models. Different models are managed by a single recognizer through their integration into a multi-model search space.

All the proposed methods were experimentally evaluated and the reduction of word error rate, from the recognizer baseline to its best configuration, is equal to 23.0%.

This paper is organized as follows. Section 2 describes the corpus utilized in the experiments. Section 3 deals with acoustic modeling, with special regard to some of the typical problems of spontaneous speech. In Section 4, the organization of the search space during recognition is shown, focusing in particular on how the integration of different sets of acoustic models (AMs) in a single network is done. Results are presented and discussed in Section 5. Conclusions and some proposals for future work end the paper.

2. CORPUS DESCRIPTION

The experiments reported below were carried out on a dialogue corpus collected at ITC-Irst [2], composed of person-to-person Italian conversations for which acoustic signals, word transcriptions and linguistic annotations are available. The two speakers were asked to fix an appointment, observing the restrictions shown on two calendar pages they were given; they did not see each other and could hear the partner only through headphones. The conversations took place in an acoustically isolated room and were naturally uttered by the speakers, without any machine mediation.

The dialogues were transcribed by annotating all extra-linguistic phenomena such as mispronunciations, restarts and human noises, with the exception of pauses.

There were 61 speakers who participated in the experiment, 22 female and 39 male, with a total of 201 dialogues.

The whole corpus was then divided into training and test sets (see Table 1), paying attention to avoid speaker overlap between the two sets. The test set consists of all the sentences uttered by 11 speakers, resulting in 20 complete dialogues and 12 half dialogues, for a total of 406 turns.

For Language Model (LM) training, 75 dialogues (601 turns) in American English, collected at CMU[1], were translated into Italian and used.

3. ACOUSTIC MODELING

Short time spectral analysis is performed over 20 *ms* Hamming windows at a rate of 10 *ms*. For every window, 12 Mel

[1] The authors would like to thank the JANUS group at CMU for giving us these dialogue transcriptions.

	Training	Test	Whole Corpus
# dialogue	169+12/2	20+12/2	201
# speaker	50	11	61
# male	32	7	39
# female	18	4	22
# turn	2680	406	3086
minutes of speech	240.4	38.2	278.6
|W| (non-noise)	27786	4683	32469
|V| (non-noise)	1291	627	1433

Frequent Human Noises			
phenomenon	# occurr.	# affected turns	%
inhalation	1509	1031	33.4
eee	941	670	21.7
exhalation	649	510	16.5
vowel lengthening inside words	642	482	15.6
mouth	515	467	15.1
mmm	211	193	6.2
ehm	113	103	3.3
ah	56	55	1.8

Table 1: Training and test set statistics.

scaled Cepstral coefficients, the log-energy and their first and second order time derivatives are evaluated.

Acoustic modeling is done using hidden Markov models (HMMs) of phone units. A set of 48 HMMs, corresponding to a subset of the SAMPA alphabet, is adopted for developing the baseline system. Models are defined by a three states left-to-right topology; the output probability densities consist of mixtures of multi-variate Gaussian functions having diagonal covariance matrices.

Baseline HMMs are initialized on a phonetically rich database, APASCI, collected at ITC-Irst [1], while a re-training phase is carried out on the herein considered spontaneous speech corpus, both in gender independent and gender dependent mode. Furthermore, the parameters of the HMMs corresponding to phones occurring few times in the training set are not re-estimated.

Extra linguistic phenomena are taken into account by training specific models. The most frequent are listed in Table 1, together with their occurrences inside the corpus.

One of the fundamental characteristics of spontaneous speech is represented by both inter- and intra-speaker variation of the speaking rate. This is observed also in our corpus where, in fact, the rate of speech often varies depending on the status of the dialogue as well as on the concepts the speaker is focusing on. It is known that there is a correlation between speaking rate variations and the presence of semantic boundaries inside dialogue turns. This means that the former could help the detection of the latter ones as shown in a paper presented in this conference [5].

A method similar to that described in [8] is used to account for speaking rate variations. Before doing re-training, the transition probabilities of HMMs are scaled to reduce the probability of staying in each state and to augment that of exit the state (see Figure 1). Hence, during the re-training phase, only the parameters of the Gaussian mixtures associated to the HMMs states are re-estimated, while transition probabilities remain unchanged. Hereafter, the HMMs thus obtained will be called "fast models".

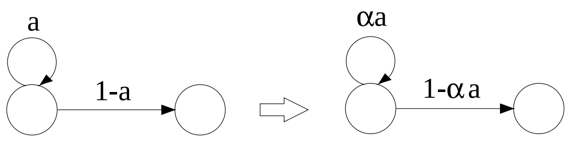

Figure 1: Transition probability scaling.

4. LANGUAGE MODELING AND SEARCH

The recognizer is derived from that developed at ITC-Irst for large vocabulary dictation tasks (20K words). The 48 context and speaker independent phonetic units are modeled by left-to-right HMMs, with three or four states. The recognizer outputs the sentence with the highest likelihood.

On the 1.3K-word vocabulary, a Shift-β trigram LM was estimated [6], including extra linguistic models as if they were real words. Intra-word acoustic constraints (phonetic transcriptions) and inter-word linguistic constraints (language model) are compiled into a sharing-tail tree-based network that defines the search space, at unit level, of the decoding algorithm [3]. Extra linguistic models are inserted into the search space by using their trigram distributions as given by the LM. Early experiments showed that this results in a (moderately) better performance than allowing their optional insertion between each pair of adjacent words.

Since little data are available for LM training, some word classes are defined such as *forename*, *place-name*, *day-name*, *month-name*, and a few classes of numbers. Moreover, some classes regarding dates and hours as combination of numbers with articles and prepositions are added.

4.1. Multi-models into the Search Space

If two or more different sets of phone models are available for a certain task, usually, given a LM for that task, a specific recognizer is built for each set of phone models [7]. For example, if gender dependent models are available, two recognizers can be implemented, one for male voices and the other for female voices. It is possible, however, to combine two (or more) recognizers, by simply defining a unique search space as the conjunction of the single ones, as shown in Figure 2.

As the search algorithm uses a beam threshold to discard less likely hypotheses, it is probable that after the processing of

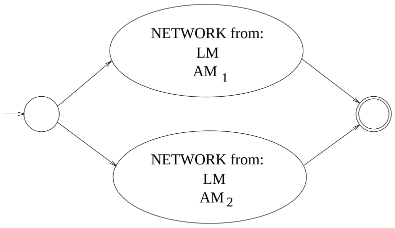

Figure 2: Multi-model search space.

the first few frames, all the surviving paths belong to the part of search space built with models that best match the input. For example, if the input voice is from a female, it is likely that all the paths in the search space, referring to male models, will soon die.

This technique is used, for example, in multilingual speech recognition [9], where the search space is defined by joining different AMs and LMs, one for each different language; as a side effect, it gives an implicit language identification. Here it is used to give the recognizer the possibility of choosing the "best" phone model set, between gender dependent and fast ones, as described in Section 5.

Note that computational time and dynamic space requirements in joining search spaces based on different models, slightly increase if models are quite different. In [9], for example, as the joined search spaces refer to different languages, the requirements for multilingual recognition approximate that for monolingual recognition.

5. EXPERIMENTS

All experiments discussed in this section were conducted on the test set reported in Table 1, while the training set was used for both acoustic re-training and LM training. Vocabulary was defined on the whole corpus, that is experiments were done with closed vocabulary by assigning to test words not occurring in the training set very low unigram probabilities.

5.1. Experimental Results

The first experiment considers the effect of adding specific models of extra linguistic phenomena into the search space. In the baseline, just a model of background noise is used, optionally introduced by the recognizer between each word pair. Then, a model for each of the five most frequent phenomena (inhalation, eee, exhalation, mouth, mmm) was trained and added to the search space as described in Section 4. As shown in Table 2, the relative reduction of Word Error Rate (WER) was equal to 17.2%. Since the improvement was remarkable, all the following experiments were done with the above mentioned extra models.

A set of experiments was then performed to assess the effectiveness of both the fast models (Section 3) and multi-model search space organization (Subsection 4.1). Results are shown in Table 3. Each row collects performance obtained by using fast models with different probability scaling factors, namely 1.0 (baseline), 0.8, 0.6, 0.4 and 0.2. Column `GenIn` gives WER obtained with gender independent models; columns `F` and `M` refer to the gender dependent models; column min(F,M) reports WER computed by choosing the best performance, sentence by sentence, between those obtained with gender dependent models. Note that the performance of min(F,M) is better than the weighted sum of F and M performance because in more than 10% of sentences the best model turns out to be that of the other gender. However, considering sentences of each speaker as a whole, the best results are always obtained with the "right" unit set (that is female speakers with female models and male speakers with male models). The last column refers to multi-model experiments, where the gender dependent models are joined in a single network, for each scale factor.

WER	extra linguistic phenomena		
	no	yes	ΔE %
baseline	33.0	27.4	17.2

Table 2: WER with or without models of extra linguistic phenomena.

α	GenIn	F	M	min(F,M)	$F\oplus M$
baseline	27.4	26.4	26.6	24.9	26.9
0.8	26.6	25.4	26.5	24.5	26.5
0.6	26.4	25.7	26.2	24.4	26.4
0.4	25.8	25.7	26.0	24.2	26.3
0.2	25.4	25.0	26.1	24.1	26.1

Table 3: WER obtained on test set using different AMs. F and M refers to the performance of male and female speakers tested on the corresponding models.

5.2. Discussion

First of all, it can be noticed that the faster the models are, the higher the performance is. In particular, the best WER on the whole test set, 25.4%, is obtained with the fastest gender-independent models, and corresponds to a relative improvement of 7.3% with respect to the unscaled models.

Although the fast models provide the best accuracy, they do not give the highest likelihood. In fact, it always happens that models scaled with a factor α produce a better likelihood, on a given test sentence, than models scaled with a factor $\alpha' < \alpha$, but the error rate is often worse. The use of the maximum likelihood paradigm during the training step justifies the fact that the baseline models give the highest likelihood. Instead, the better performance of fast models can be explained with the larger amount of "fast" speech than "slow" speech in the test material.

A further analysis was carried out with the purpose of evaluating the ideal system performance by choosing the best hypothesized sentence between those obtained using the five differently scaled gender-independent model sets. The resulting WER is 22.3%, that is, a relative WER reduction of 18.6% (with respect to the baseline system) could be obtained if we were able to predict the best scale factor for each sentence. However, in this case, the multi-model search space organization for joining the five different AMs could not be adopted, since the best likelihood would be always given by the baseline model set, as stated above.

As shown in the last column of the table, joining gender dependent models improves performance over gender independent ones only when no scaling of HMM transition probabilities is done, even if the potential improvement (see min(F,M) column) is relevant. For completeness, the WER obtained by choosing, sentence by sentence, the best performance between those of differently scaled gender-dependent models is 22.0%.

Improvements could be expected if the scale factor for HMM transition probabilities were dynamically changed during decoding, according to an estimation of the current speaking rate. Some work related to this problem can be found in [8].

To better model the duration of each phone unit, enlarging the number of HMM states and introducing skips among states was tried. Furthermore, experiments using a scale factor dependent on some predefined phonetic classes (e.g. vowels, stops, fricatives, nasals, liquids) were conducted. However, in both cases no improvements were observed.

6. CONCLUSIONS AND FUTURE WORK

An ASR system for spontaneous speech dialogues was described and some methods for increasing performance with respect to a HMM-based system were proposed. These methods consist in the introduction of specific models of extra-linguistic phenomena and in the scaling of transition probabilities of HMMs for coping with speaking rate variations. Both gender independent and gender dependent models were used, combining some of them into a multi-model search space.

To evaluate all the proposed methods, experiments were carried out on a corpus of spontaneously uttered dialogues, and results were presented.

The introduction of specific spontaneous speech models allows a considerable improvement in performance. Also the use of models adapted to fast speech is advisable. In this last case an estimate of the speaking rate parameter during the course of the dialogue could lead to further improvements. A method for deriving such estimate is proposed in [5]. Furthermore, the use of the spectral variation functions defined in [4] could be helpful for this task.

7. REFERENCES

1. B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1391–1394, Yokohama, Japan, 1994.
2. B. Angelini, M. Cettolo, A. Corazza, D. Falavigna, and G. Lazzari. Multilingual Person to Person Communication at IRST. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997.
3. F. Brugnara and M. Cettolo. Improvements in Tree-based Language Model Representation. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1797–1800, Madrid, Spain, 1995.
4. F. Brugnara, D. Falavigna, and M. Omologo. Automatic Segmentation and Labeling of Speech based on Hidden Markov Models. *Speech Communication*, 12:357–370, 1993.
5. M. Cettolo and D. Falavigna. Automatic Detection of Semantic Boundaries based on Acoustic Lexical Knowledge. In *Proceedings of the International Conference on Spoken Language Processing*, Sidney, Australia, 1998.
6. M. Federico, M. Cettolo, F. Brugnara, and G. Antoniol. Language Modeling for Efficient Beam-Search. *Computer Speech and Language*, 9(4):353–379, 1995.
7. L.F. Lamel and J.L. Gauvain. High Performance Speaker-Independent Phone Recognition Using CDHMM. In *Proceedings of the European Conference on Speech Communication and Technology*, pages I:121–124, Berlin, Germany, 1993.
8. N. Mirghafori, E. Folser, and N. Morgan. Towards Robustness to Fast Speech in ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages I:335–338, Atlanta, Georgia, USA, 1996.
9. E. Noeth, S. Harbeck, and H. Niemann. Multilingual Speech Recognition. In *Informal Proceedings of the NATO ASI school on "Computational Models of Speech Pattern Processing"*, St. Helier, Jersey, UK, 1997. To be published by Springer Verlag.