Acoustics'08 Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Speech transcription for Embodied Conversational Agent animation

Leila Zouari and Gerard Chollet

Telecom Paris Tech, 46 rue Barrault, 75013 Paris, France
zouari@enst.fr

Abstract

This article investigates speech transcription within a framework of Embodied Conversational Agent (ECA) animation by voice. The idea is to detect some pronounced expressions/keywords in order to animate automatically the face and the body of an avatar.

Extensibility, speed and precision are the main constraints of this interactive application. So after defining the set of the relevant words (to the application), a fast large vocabulary speech recognition system was developed and the keyword detection was evaluated.

In order to fasten the recognition system without decreasing its efficiency, the acoustic models have been shortened by an original process. It consists in decreasing the number of shared central states of context dependent models which are considered stationary. The shared states located in the border of the models remain inchanged. Then all the models are retrained.

The system is evaluated on an hour of the ESTER database (a French broadcast news database). The experiments show that reducing the number central states of triphones is advantageous. Indeed, the length of models is reduced by 20% with no loss of accuracy.

1 Introduction

The widespread of dialog systems in daily services has created a great demand for user friendly interfaces. In particular, Embodied Conversational Agents (ECA) systems are computer interfaces represented by lifelike human or animal characters. These systems may allow humans to dialog with machines naturally and easily.

In order to make the ECA able to perform believable actions when they communicate with human users they must be animated according to what they are saying.

As ECA are often employed by interactive applications, speech-based dialog must be operated in real time. Hence it is necessary to have fast and accurate speech recognition systems.

State of the art speech recognition systems are based on Hidden Markov Models (HMM). Often, the states of an HMM are modeled by mixtures of Gaussian distributions (GMM). As the performance and the speed of speech recognition systems are closely related to the number of HMM/GMM Gaussians, reducing the number of Gaussians without decreasing the system performance is of major interest. Mainly a trade off is to be found between the model precision and the ability to accurately estimate the model parameters.

For this reason many fast GMM computation techniques were proposed in the literature [1, 5, 7, 6, 9, 8, 2]. These methods can be organized into four levels [3] : the frame level, the GMM level, the Gaussian level and the component level.

In this paper we propose a method belonging to the Gaussian level. This method operates as follows :

1. Context dependent models with tied states are created and estimated on the training data.
2. For each phone, the states located in the middle of each allophone are tied.
3. The obtained models are re-estimated.

This proposed algorithm is evaluated within the framework of a large vocabulary continuous speech recognition task. The results on ESTER, a French broadcast news database, show that the length of models is reduced by 20% with no loss of accuracy.

The remainder of this paper is organized as follows: section 2 outlines the context of this work which is the development of a multi modal human machine interface, section 3 describes the proposed algorithm for models shortening, section 4 reports on tests protocols and results, section 5 outlines the conclusions and the prospective work.

2 Human Machine Interaction

A part of the work described below have been conducted within the framework of the project MyBlog3D (https://picoforge.int-evry.fr/cgi-bin/twiki/view/Myblog3d/Web/). The goal of this project is to build a multi modal human-machine dialog system for a virtual and augmented reality application.

When somebody faces a camera, he is represented by an avatar in a virtual world. The avatar have the same face and voice as the real person. The face and the body of the avatar are animated according to the speech of that person. In order to hide his identity, the person can change the voice and the face of the avatar.

This muti modal application involves many technologies belonging to different fields, which are mainly speech processing (recognition, conversion and synthesis), Embodied Conversational Agents (ECA), and gesture analysis and synthesis.

In this paper, we are interested by only the speech part. The global architecture of this system is described in figure 2.

3 Models shortening

State tying is a common procedure, generally used by speech recognition systems, to reduce the models complexity. This way, they can be re-estimated by means of only a limited amount of data.

Due to the co-articulation phenomena, the states in the border of a hidden Markov models are influenced by the previous or the following phones. The central state can be considered stationary.

So, the idea proposed in this work is to limit the number of central states in order to reduce the total number of states without decreasing the system performance.

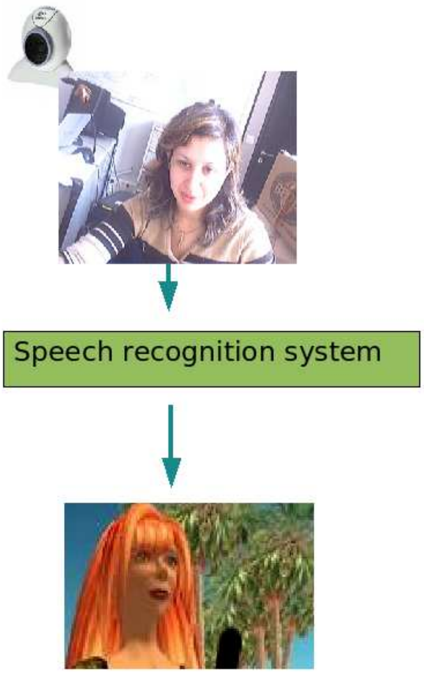

Figure 1: The system architecture

This algorithm performs as follows :

- Context dependent models (generally triphones or quinphones) are created and their parameters are estimated on the training data.
- For each phone, the states in the border of its allophones are tied according to the classical decision tree based procedure [11].
- Central states are tied into a fixed number that is we stop the tree growing when the proposed number of states is reached.

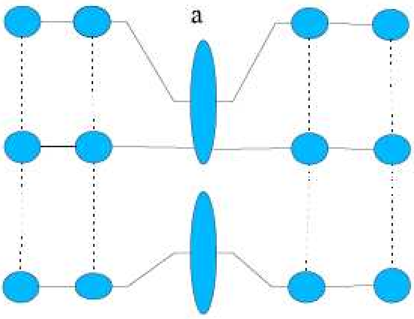

Figure 2: Example of central state tying for the allpohones of the phone "a"

The difference between the proposed procedure and the standard state tying one [11] is that in the first case the number of central states is limited by a fixed number. The states in the model's border are not affected by this procedure.

4 Experiments and results

4.1 Experimental setup

A large vocabulary speech recognition system based on Sphinx training and test tools (SphinxTrain and Sphinx3.0.5 [10]) is developed.
The resources (the acoustics models, the language models and the dictionary) are created and evaluated on the ESTER database. ESTER [4] is a French broadcast news database. It contains 100 hours of audio and 400 millions of written words (text data).
The audio data is transcribed manually. The resulting transcription contains extra information such as the speaker gender and/or identity, the recording conditions, ...
This data is divided into 3 parts : 82 hours for training, 8 hours for the development and 10 hours for the evaluation (see table 7).

Source	Train/Dev	Test
France Inter	32h20 / 2h	2h
France info	8h/2h	2h
RFI	23h/2h	2h
RTM	18h/2h	2h
Radio classique	-	1h
Culture	-	1h
Total	90h	10h
Période	1998-2000/2003	2004

Table 1: Audio data

The text data comes mainly from the newspaper "Le Monde" (from 1987 to 2003).

4.2 Speech recognition

Many steps are required to develop a speech recognition system : parameter extraction, acoustic and linguistic resources preparation, decoding and evaluation.

4.2.1 Parameter extraction

The parameter vectors are composed of 12 MFCC (Mel frequency Cepstral coefficients) coefficients, energy, and their first and second derivatives. For each sentence, these vectors are normalized with respect to the mean and the variance in order to improve the system robustness against the speaker changes and the recording conditions.

4.2.2 Acoustic models

Two kinds of acoustic models are developed. The first models are context independent with 32, 64, .., 512 Gaussians per state. The second models are cross-word context dependent with 6108 tied states and 32 or 16 Gaussians per state. The parameters of these models are estimated on the Ester train database.

4.2.3 Dictionary

The dictionary contains 65000 distinct words with one to eight pronunciations for each, that is a total of 118000 words.

4.2.4 Language modeling

The language model is trigram with 16 millions of trigrams and 17 millions of bigrams. Audio and text data (coming from the ESTER database and the newspaper "Le Monde") are used to construct this model.

4.2.5 Experimental results

All the experiences have been conducted on the same standard personal computer. This computer have the following characteristics : processor 3.6 Ghz, 6 Go of RAM, 50Go of hard disc and 512 Kb of cache memory.

The monophones are first evaluated on only one hour of "Radio classique". The results are as follows :

Gaussians number	WER (%)
32	37.7
64	36.7
128	35.6
256	35.2
512	35.3

Table 2: Recognition results of monophones on "Radio classique"

WER is the Word Error rate.

The performance of the monophones are not good enough, so we experimented the triphones.
The triphones have been evaluated on each radio station and the on all the database test set (see table 3).

Radio	WER(%)
Classique	28.7
Culture	40.2
Info	31.5
Inter	35.6
RFI	36.4
RTM	43.8
Total	36.2

Table 3: Recognition results of triphones on ESTER test set

We can see that the triphones are more performant than the monophones which is foreseeable. In addition, the results for the different radio stations are vary different. The triphones based system operates in three times the real time.

4.3 Keyword spotting

4.3.1 Keyword selection

As this work is done in the framework of the project MyBlog3D, the keyword detection is operated on the words/expressions fixed by the project application.

The first step is then to find the relevant words for the application and verify that they belong to ESTER database. The results of these studies are illustrated by the tables 4 (for the train part) and 5 (for the test part).

Keywords	Train
bonjour	825
tout le monde	212
au revoir	17
gauche	203
droite	245
haut	394
bas	1004
d'accord	155
Total	3055

Table 4: Keywords presence in the train set of the ESTER database

Keywords	Test	R. Classique
bonjour	60	11
tout le monde	27	2
au revoir	2	1
gauche	14	1
droite	11	0
haut	26	1
bas	92	4
d'accord	17	3
Total	249	23

Table 5: Keyword presence in the test set of ESTER and in "Radio Classique"

4.3.2 Evaluation on one hour

This test concern the evaluation of the keyword detection on the "Radio Classique" station. The results are reported as follows :

Words	Sentences	Precision (%)
837	9424	70.1

Table 6: Keyword detection results on "Radio Classique"

These results are not significant because the evaluated keywords are not present enough in the part of the test set "Radio Classique".

4.3.3 Evaluation on 10 hours

The same evaluation is now conducted on all the database test set (10 hours). The results are the following :

Words	Sentences	Precision (%)
11014	108717	82.8

Table 7: Keyword detection results on Ester test set

4.4 Model shortening

The model shortening algorithm (described in section 3) have been applied on two kinds of models : triphones wth 16 Gaussians per state and triphones with 32 Gaussians per state. Tables 8 are 9 report the corresponding results.

Number of central states	WER(%)
2	26.9
4	27.1
8	27.2
initial models	27.6

Table 8: 16

For the models with 16 Gaussians per state, reducing the number of central states is advantageous. In fact, with only two states, the length of the model is reduced and the WER decreases by about 0.7% absolute.

Number of central states	WER(%)
2	27
4	26.9
8	26.7
initial system	27.1

Table 9:

For the models with 32 Gaussians per state, reducing the number of central states is also advantageous. But when this number is too limited, the WER begins to increase. The reason is that the resulting number of states is not sufficient to estimate the models parameters. For the best configuration, the number of total states is reduced to 5000 (by 20%) with no loss of accuracy.

5 Conclusion

This paper presents an algorithm of acoustic models shortening to fasten a large vocabulary speech recognition system. Such system, designed for a multi modal interactive dialog in a virtual word, must be real time operating.
For each phone, we keep the same number of tied states in the allophones border but we reduce the central states by a further tying.
Experiments on the Ester broadcast news database show that reducing the number of central tied states is advantageous. In fact the length of models is reduced by 20% with no loss of accuracy.

6 Acknowledgment

This work is performed in the framework of the project ANR-MyBlog3D.
The authors would like to thank Kader Oukaci for his participation to the experiences.

References

[1] E. Bocchieri. Vector Quantization for the Efficient Computation of Continuous Density Likelihoods. In *International Conference on Acoustics, Speech, and Signal Processing ICASSP*, volume 2, pages 692–695, Minneapolis, Avril 1993.

[2] E. Bocchieri and B. Mak. Subspace Distribution Clustering Hidden Markov Model. *IEEE transactions on Speech and Audio Processing*, pages 264–275, Mars 2001.

[3] A. Chan, J. Sherwani, R. Mosur, and A. Rudnicky. Four Layer Categorization Scheme of Fast GMM Computation Techniques in Large Vocabulary Continuous Speech Recognition Systems. In *proceedings ICSLP*, Jesu Island - Korea, 2004.

[4] G. Gravier, JF. Bonastre, S. Galliano, E. Geoffrois, K. Mc-Tait, and K. Choukri. ESTER Une Campagne d'Evaluation des Systčmes d'Indexation d'Emissions Radiophoniques. In *Journées d'Etude sur la Parole JEP*, pages 253–256, Fčs, Avril 2004.

[5] KM. Knill, MJF. Gales, and S. Young. State based Gaussian Selection in Large Vocabulary Continuous Speech Recognition using HMMs. In *IEEE Transactions on Speech and Audio Processing*, volume 7, pages 152–161, Mars 1999.

[6] J. Leppänen and I. Kiss. Gaussian Selection with Non-overlapping Clusters for ASR in Embedded Devices. In *International Conference on Acoustics, Speech, and Signal Processing ICASSP*, France, May 2006.

[7] J. Olsen. Gaussian Selection using Multiple Quantisation Indexes. In *IEEE Nordic Processing symposium*, 2000.

[8] S. Ortmanns, H. Ney, and T. Firslaff. Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition. In *European Conference on Speech Communication and Technology*, pages 139–142, Rhodčs, Septembre 1998.

[9] M. Padmanablan, L. Bahl, and D. Nahamoo. Partitioning the Feature Space of a Classifier with Linear Hyperplanes. In *IEEE Transactions on Speech and Audio Processing*, volume 7, pages 282–288, May 1999.

[10] R. Singh. Sphinxtrain. http://fife.speech.cs.cmu.edu/sphinxman, Novembre 2000.

[11] S. Young, J. Odell, and P. Woodland. Tree-based State Tying for High Accuracy Acoustic Modeling. In *proceedings ARPA Workshop on Human Language Technology*, pages 307–312, 1994.