PROGRESSES IN CONTINUOUS SPEECH RECOGNITION BASED ON STATISTICAL MODELLING FOR ROMANIAN LANGUAGE

Corneliu Octavian Dumitru[1,2]
[1]*University Politehnica Bucharest, Faculty of Electronics Telecommunications and Information Technology, Romania*
[2]*ARTEMIS Department, GET/INT, Evry, France*

Inge Gavat, Diana Militaru
University Politehnica Bucharest, Faculty of Electronics Telecommunications and Information Technology, Romania

Keywords: MFCC, LPC, PLP, statistical modelling, monophone, triphone.

Abstract: In this paper we will present progresses made in Automatic Speech Recognition (ASR) for Romanian language based on statistical modelling with hidden Markov models (HMMs). The progresses concern enhancement of modelling by taking into account the context in form of triphones, improvement of speaker independence by applying a gender specific training and enlargement of the feature categories used to describe speech sequences derived not only from perceptual cepstral analysis but also from perceptual linear prediction.

1 INTRODUCTION

After years of research and development, the problem of automatic speech recognition and understanding is still an open issue. The end goal of translation into text, accurate and efficient, unaffected by speaker, environment or equipment used is very difficult to achieve and many challenges are to be faced.

Some factors which make difficult the problem of automatic speech recognition (ASR) are voice variations in context or in environment, syntax, the size of vocabulary. How this problems can easy be accommodated by humans, to continue studies in human speech perception can be very important to improve performance in speech recognition by machine. As alternative, human performance can be regarded as guide for ASRs and in this moment neither the best systems built for English language can not reach human performance.

In this paper we will present progresses made in ASR for Romanian language based on statistical modelling with hidden Markov models (HMMs) using the HTK toolkit, developed by the Cambridge University Engineering Department (Woodland, 1994). We started in this domain with a system for continuous speech recognition (Gavat, 2003) based on monophone models, regarding words like phone sequences, neglecting co-articulation. The obtained results challenged us to improvements that addressed especially acoustical context modelling but also speaker independence enhancement and optimization of features choice describing speech.

The remainder of this paper will be also structured as follows: Section 2 presents an overview of the built ASR system. Section 3 describes the context based acoustical modelling, realized with triphones. Section 4 gives an overview of the Rumanian database and the changes made in view of the experiments for speaker independence enhancement. Tests are described and discussed in section 5. Finally, conclusions and future works are given in section 6.

2 SYSTEM OVERWIEV

The functional schema of our speech recognition system is given in figure 1. The system acts in three

Octavian Dumitru C., Gavat I. and Militaru D. (2007).
PROGRESSES IN CONTINUOUS SPEECH RECOGNITION BASED ON STATISTICAL MODELLING FOR ROMANIAN LANGUAGE.
In *Proceedings of the Fourth International Conference on Informatics in Control, Automation and Robotics*, pages 262-267
DOI: 10.5220/0001619802620267

main phases: training, testing and evaluation of results.

In the training phase the current parameters of HMMs are established by incrementally refining an initial set of "white" acoustical HMM models.

In the testing phase the test data are verified with the training database using the grammar, the lexicon, the word network and the testing dictionary.

In the evaluation phase, by comparing the transcription of the test speech sequences with the reference transcription, the recognition correctness and accuracy are calculated.

In each phase a data preparation stage is introduced in order to build a set of speech data files and their transcription in the required format. An important step in this stage is feature extraction.

Feature extraction is the lowest level of automatic speech recognition and it lies in the task of extracting the limited amount of useful information from high-dimensional data. The feature extractors maintain much of the characteristics of the original speech and eliminate much of the extraneous information.

After feature extraction, a sequence O of feature vectors are the input data for the testing phase, in order to establish the correct uttered sentence W. We need to estimate therefore the probability of acoustic features given the phonetic model, so that we can recognize the input data for the correct sentence. This probability is referred to as acoustic probability P (O|W).

HMMs are the most common models used in automatic speech recognition systems to model the joint probability distribution of feature vectors for a given utterance model (Gavat, 2000).

Mathematically the problem in continuous speech recognition is to find a sequence of words $\hat{W}$ such that

$$\hat{W} = \arg\max_{W} P(O|W)P(W) \quad (1)$$

The most probable sentence W given the observation sequence O can be computed by taking this product of two probabilities for each sentence and choosing the sentence for which the product is the highest.

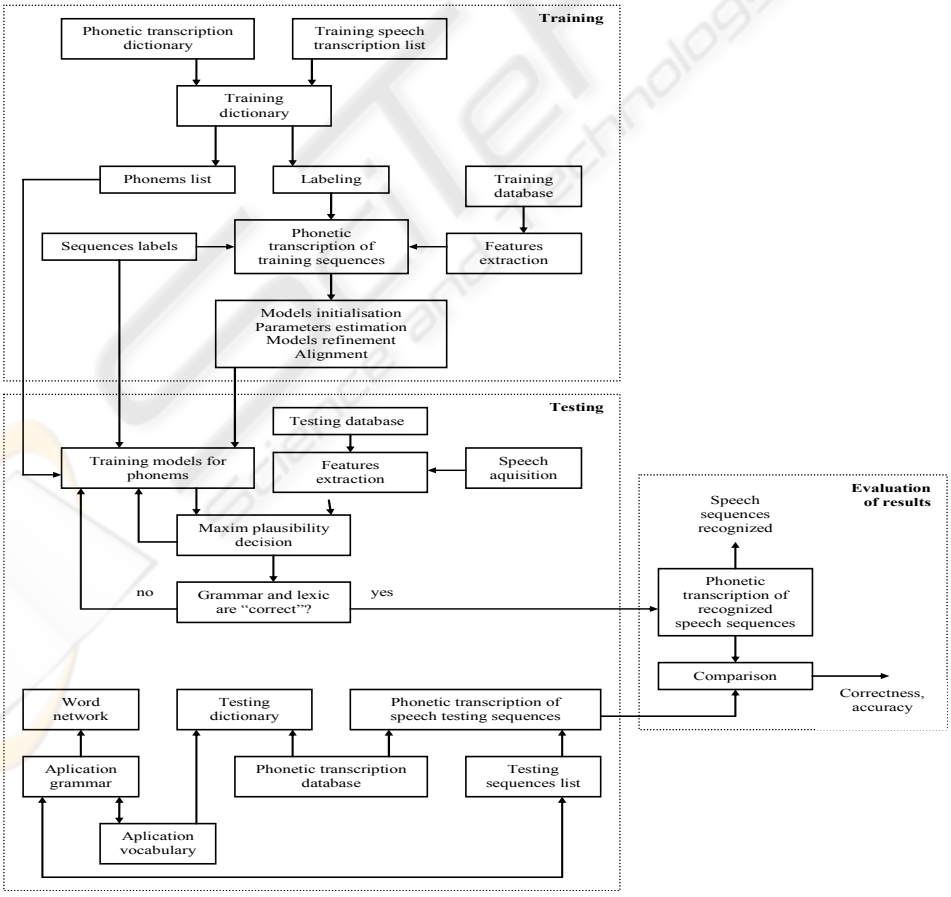


Figure 1: The speech recognition system.

The prior probability P (W) is calculated using a language model appropriate for the recognition task, and the acoustic probability P (O|W), is calculated by concatenating the HMMs of the words in the sequence W and using the Viterbi algorithm for decoding. A silence or a 'short pause' model is usually inserted between the HMMs to be concatenated.

3 TRIPHONE MODELLING

For small vocabulary recognition, word models are widely used, since they are accurate and trainable. In the situation of a specific and limited task they become valid if enough training data are available, but they are typically not generalizable. Therefore, usually for not very limited tasks are preferred phonetic models based on monophones, because the phones, as smallest linguistic units, are easy generalizable and of course also trainable.

Monophones constitute the foundation of any training method, in any language, and we also started with them. But a refinement of this initial step was necessary because in real speech, the words are not simple strings of independent phonemes. The realization of a phoneme is strongly affected by its immediately neighboring phonemes by co-articulation. Because of this, monophone models have been changed in time with triphone models that became the actual state of the art in automatic speech recognition with large vocabularies (Young, 1992), (Young, 1994).

A triphone model is a phonetic model that takes into consideration the left and the right neighbouring phones. This immediate neighbour – phonemes are called respectively the left and the right context; a phoneme constitutes with the left and right context a triphone. For example in the SAMPA (Speech Assessment Methods Phonetic Alphabet) transcription "m - a + j" of the Romanian word "mai", regarded as triphone, the phoneme "m" has as left context "a" and as right context "j".

For each such a triphone a model must be trained: in Romanian that will give a number which equals 40,000 models, situation totally unacceptable for a real world system. In our speech recognition task we have modelled only internal – word triphones and the adopted state tying procedure has conducted to a controllable situation.

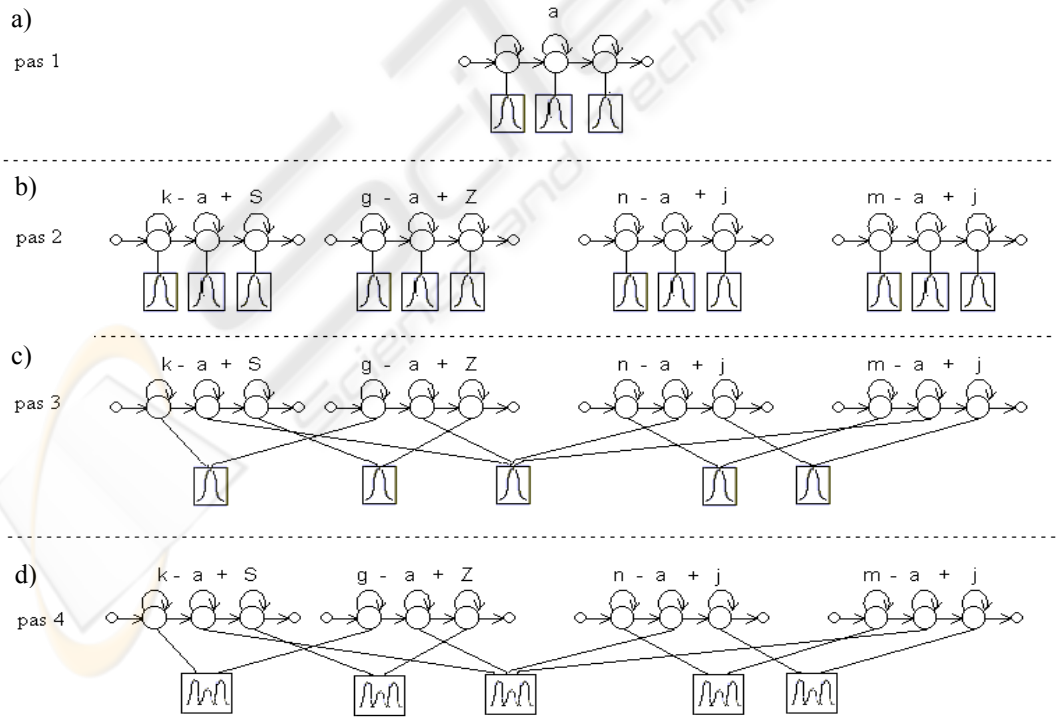


Figure 2: Different models for triphones around the phoneme "a".

If triphones are used in place of monophonemes, the number of needed model increases and it may occur the problem of insufficient training data. To solve this problem, tying of acoustically similar states of the models built for triphones corresponding to each context is an efficient solution. For example, in figure 2b, four models are represented for different contexts of the phoneme "a", namely the triphones "k – a + S", "g – a + z", "n – a + j", "m – a + j". In figure 2c, 2d, there are represented the clusters formed with acoustically similar states of the corresponding HMMs.

The choice of the states and the clustering in phonetic classes are achieved by mean of phonetic decision trees. A phonetic decision tree built as a binary tree, is shown in figure 3 and has in the root node all the training frames to be tied, in other words all the contexts of a phoneme. To each node of the tree, beginning with the parent – nodes, a question is associated concerning the contexts of the phoneme. Possible questions are, for example: is the right context a vowel (R = Consonant?), is the left context a phoneme "a" (L = a?); the first answer designates a large class of phonemes, the second only a single phonetic element. Depending on the answer, yes or no, child nodes are created and the frames are placed in them. New questions are further made for the child nodes, and the frames are divided again.

The questions are chosen in order to increase the log likelihood of the data after splitting. Splitting is stopped when increasing in log likelihood is less than an imposed threshold resulting a leaf node. In such leaf nodes are concentrated all states having the same answer to the question made along the path from the root node and therefore states reaching the same leaf node can be tied as regarded acoustically similar. For each leaf node pair the occupancy must be calculated in order to merge insufficient occupied leaf nodes.

A decision tree is built for each state of each phoneme. The sequential top down construction of the decision trees was realized automatically, with an algorithm selecting the questions to be answered from a large set of 130 questions, established after knowledge about phonetic rules for Romanian language.

4 DATABASE

The data are sampled by 16 kHz, quantified with 16 bits, and recorded in a laboratory environment.

For continuous speech recognition, database for training is constituted by 3300 phrases, uttered by 11 speakers, 7 males and 4 females, each speaker reading 300 phrases.

The databases for testing contained 220 phrases uttered by 11 speakers, each of them reading 20 phrases.

The training database contains over 3200 distinct words; the testing database contains 900 distinct words.

In order to carry out our experiments about speaker independence, the database was reorganized as follows: one database for male speakers (MS), one database for female speakers (FS) and one database for male and female speakers (MS and FS). In all cases we have excluded one MS and one FS from the training and used for testing.

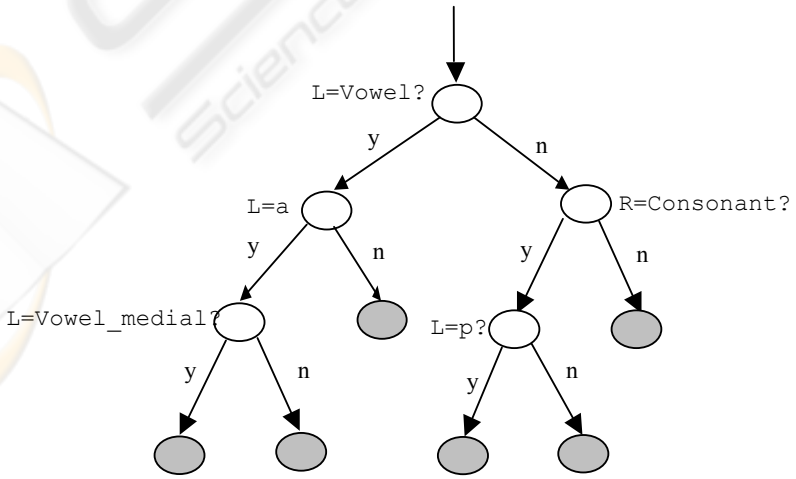


Figure 3: Phonetic tree for phoneme m in state 2.

The speech files from these databases were analysed in order to extract the interesting features. The feature extraction methods used are based on linear predictive coding (LPC), perceptual linear prediction (PLP) (Hermansky, 1990) mel-frequency cepstral coefficients (MFCC).

5 EXPERIMENTAL RESULTS

To assess the progresses made with our ASR system we initiated comparative tests for the performance expressed in word recognition rate (WRR) to establish the values under the new conditions versus the preceding ones. The comparison is made for the following situations:

- Triphone modelling/monophone modelling
- Gender based training/mixed training
- LPC and PLP/MFCC.

The results obtained in the experiments realized under these conditions are summarized in Table 1, Table 2, and Table 3.

Table 1: Word Recognition Rate: Training MS, testing MS or FS.

Training MS	Type	WRR (%)		
		MFCC_ D_A	LPC	PLP
Testing MS	Monophone	56.33	30.85	34.02
	Triphone	81.02	49.73	68.10
Testing FS	Monophone	40.98	23.23	25.12
	Triphone	72.86	47.68	59.00

The WRR are:

- For 12 LPC coefficients the word recognition rates are low: 30.85% (monophone) training and testing with MS and 49.73% (triphone); 31.11% (monophone) training and testing with FS and 61.15% (triphone); 26.10% (monophone) training MS and FS and testing with MS and 51.5% (triphone).
- For 5 PLP coefficients the obtained results are very promising, giving word recognition rates about 58.55% (triphone training and testing FS), 68.10% (triphone training and testing MS) and 70.11% (triphone training MS and FS and testing MS).
- For 36 MFCC_D_A coefficients (mel-cepstral coefficients with first and second order variation) we obtained the best results, as we expected: monophone 56.33% and triphone 81.02%, training and testing with MS; monophone 56.67% and triphone 78.43%, training and testing with FS; monophone 57.44% and triphone 78.24%, training MS and FS and testing with MS.

Table 2: Word Recognition Rate: Training FS, testing MS or FS.

Training FS	Type	WRR (%)		
		MFCC_ D_A	LPC	PLP
Testing MS	Monophone	53.56	26.72	23.78
	Triphone	69.23	49.73	53.02
Testing FS	Monophone	56.67	31.11	34.22
	Triphone	78.43	61.15	58.55

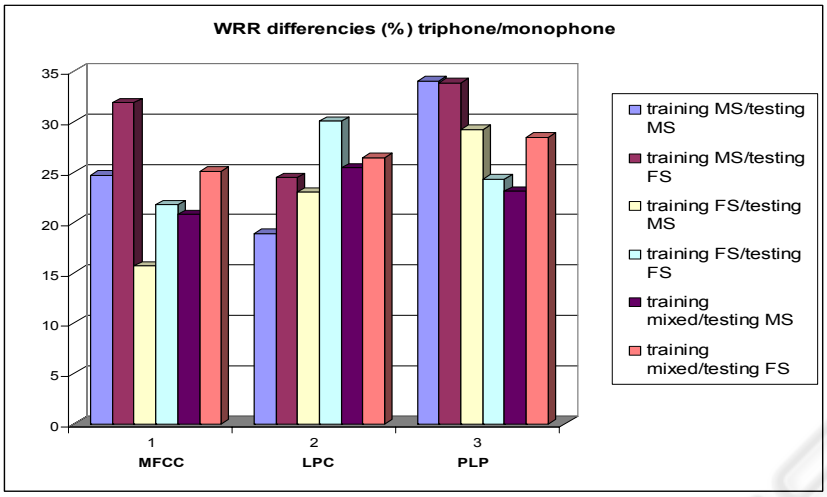
Table 3: Word Recognition Rate: Training MS and FS, testing MS or FS.

Training MS and FS	Type	WRR (%)		
		MFCC_ D_A	LPC	PLP
Testing MS	Monophone	57.44	26.10	47
	Triphone	78.24	51.50	70.11
Testing FS	Monophone	49.89	24.06	41.22
	Triphone	74.95	50.49	69.65

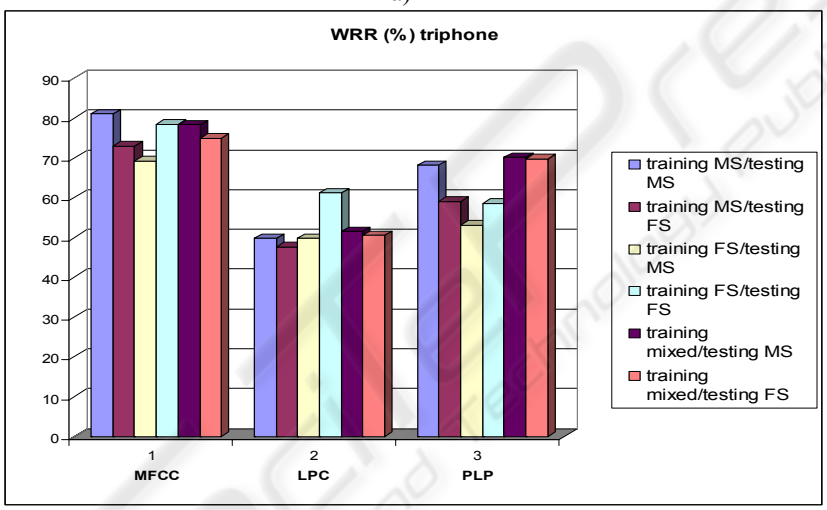
6 CONCLUSION

After the experiments made we have following conclusions:

- The triphone modelling is effective, conducting to increasing in WRR between 15% and 30% versus the monophone modelling. The maximal enhancement exceeds 30% for training MS and testing FS for MFCC_D_A (see figure 4.a).
- A gender based training conduct to good result for test made with speakers from the same gender (training MS / testing MS: 81.02%, testing FS: 72.86%; training FS/testing FS: 78.43%, testing MS: 69.23%); changing gender in testing versus training leads to a decrease in WRR around 10%. For a mixed trained data base changing gender determines only variations around 5% in WRR (see figure 4.b).



a)



b)

Figure 4: a) Chart 1; b) Chart 2.

REFERENCES

Woodland P.C., Odell J.J., Valtchev V., Young S.J, 1994, Large Vocabulary Continuous Speech Recognition Using HTK, *Proceedings of ICASSP 1994,* Adelaide.

Gavat I., Dumitru C.O., Costache G., Militaru D., 2003, Continuous Speech Recognition Based on Statistical Methods, *Proceedings of SPED 2003*, pp. 115-126, Romania.

Young, S.J., 1992, The General Use of Tying in Phoneme-Based HMM Speech Recognizers, *Proceedings of ICASSP 1992*, vol. 1, pp. 569-572.

Young S.J., Odell J.J., Woodland P.C., 1994, Tree Based State Tying for High Accuracy Modelling, *ARPA Workshop on Human Language Technology*.

SAMPA (Speech Assessment Methods Phonetic Alphabet), http://www.phon.ucl.ac.uk/home/sampa/home.htm

Hermansky H., 1990, Perceptual Linear Predictive Analysis of Speech, *J. Acoust. Soc. America*, vol.87, no.4, pp. 1738-1752.

Oancea E., Gavat I., Dumitru C.O., Munteanu D., 2004, Continuous speech recognition for Romanian language based on context-dependent modelling, *Proceedings of COMMUNICATION 2004*, pp. 221-224, Romania.

Dumitru C.O., Gavat I., 2005, Features Extraction, Modelling and Training Strategies in Continuous Speech Recognition for Romanian Language, *Proc. EUROCON 2005*, pp. 1425-1428, Serbia & Montenegro.

Gavat I., Zirra, M., Grigore O., Sabac B, Valsan Z., Cula O., Pascu A., 2000, *Elemente de sinteza si recunoasterea vorbirii*, Ed. Printech, Bucharest, Romania.

Lupu, E., Pop, G. Petre, 2004, *Prelucrarea numerica a semnalului vocal. Elemente de analiza si recunoastere*, Ed. Risoprint , Cluj-Napoca, Romania.