Speech Segmentation to Phonemes Based on Hybrid Hidden Markov Models

Yan Jingbin, Wu Shi, Kheidorov I.E., Tkachenia A.V.
Belarusian State University, Nezalezhnosti av., 4. 220050, Minsk, Belarus,
E-mail: igorhmm@mail.ru, Telephone: office +375 29 6210031, mobile +375 29 8703533

Abstract: *In this paper we develop automatic speech segmentation to phonemes using hybrid system based on Hidden Markov Model (HMM) and Artificial Neutral Network (ANN). It was shown that usage of ANN in order to estimate local probability in HMM leads to optimal global probability estimation in the general case, without imposition of additional model conditions. The result of automatic segmentation is close to the manual one, and can be successfully used in real applications for speech data segmentation and speech recognition system training.*

Keywords**:** Speech Segmentation, Hybrid Systems, Hidden Markov Model, Artificial Neutral Network.

1. INTRODUCTION

During the last 20 years stochastic finite state machine and different modifications of HMM have been successfully used for solving different recognition tasks, like speech recognition, time sequence prediction, biomedical signals analysis and others. Presently the task of speech segmentation to phonemes is one of the actual tasks due to a series of reasons. First of all, methods of speech segmentation are very import for the development of continuous speech recognition systems as acoustical probability estimation method. Secondly, accurate phoneme segmentation algorithms can be used for speech search and indexing systems in multimedia data bases.

Speech segmentation task can be defined in the following way: it is necessary to divide speech on fragments which belong to different phonemes the utterance consists of. In the general case the development of segmentation system must not have any restrictions, i.e. it has not to be sensitive to language, sex or other speaker limitations (speaker with vocal track pathologies and other specific). This segmentation task has to realize common task criterion and use speaker independent methods. The result of our work is the estimation of segments boundaries called rough boundaries. In addition the classification has to be done, i.e. the phonetic description of each segment has to be known.

For speech segmentation task single HMM is assigned to each phoneme. During training procedure a new HMM is created for each training sentence. HMM is a union of proper models. Model parameters are changed for fulfillment of probability criterion. In the last year some laboratories from all over the world announced speaker independent continuous speech recognition systems with large vocabulary, based on HMM. HMM is very good for description speech time aspects (scaling time) and insensitive to frequency distortions. HMM has many effective high-capacity training and decoding algorithms [1, 2, 3, 4]. Only sentence phonetic transcription is needed to train the system, and segmentation of training data is unnecessary. Also HMM structure can be easily increased for taking into account phonological and syntactical rules. But assumptions, which is made HMM effective and easy optimize, limited its generality. As result, HMM has a number of disadvantages [5]. For overcoming some of these drawbacks, a lot of scientists integrate ANN into HMM.

2. HYBRID SYSTEMS BASED ON HMM AND ANN

The idea to join HMM and ANN was motivated that HMM and ANN are mutually complementary features: HMM is good to used for sequential data, but some assumptions limited its generality; ANN can approximate any nonlinear function, be very flexible and not have strict assumptions about input data distribution, but ANN can not correctly process time sequences. Therefore a lot of hybrid models have been suggested and developed by all scientists.

Hybrid HMM/ANN systems. HMM is based on strict probabilistic formalism, which is made them difficult to integrate with parts of heterogeneous systems. But in paper [1, 6] it was shown that if every output ANN element is associated with state k of state sequence $\theta = \{1,2,\ldots,K\}$, it is possible to train ANN for generating a good estimated value of posterior probability of output classes. In other words, if $g_k(x_t \mid \Theta)$ is the output function k of ANN, and x_t – observation vector, that $g_k(x_t \mid \Theta^*) \approx p(q_t = k \mid x_t)$, where Θ^* – the best set of ANN parameters.

Using posterior probability (instead of local probability) in finite state machine, model becomes a recognition model, where the observation sequence is the system input and all local and global measures are based on posterior probabilities. For accommodation of this formalism it is necessary to review basis of a stochastic finite state machines. It can be shown that $p(M \mid X, \Theta)$ can be presented in terms of *transition conditional probability* $p(q_t \mid x_t, q_{t-1})$ and optimal ANN parameters Θ can be trained according to *maximum posterior probability* (MPP):

$$\Theta^* = \arg\max_{\Theta} p(M \mid X, \Theta) = \arg\max_{\Theta} p(X \mid M, \Theta)\, p(M \mid \Theta), \quad (1)$$

Total training algorithm is called recursive estimate criterion and posterior probability maximization (REMAP) and it is a training algorithm based on expectation maximization criterion (EM) which directly includes posterior probabilities and estimates desired target ANN distribution at the previous stage. Because this expectation maximization procedure has iterative

maximization stage, so it is often called the generalize expectation maximization.

The other popular scheme, when hybrid HMM/ANN systems are used as sequences recognition models, is based on change of local posterior probability to scaled probability using division posterior probability by class probability model estimation, i.e.:

$$\frac{p(q_t = k \mid x_t)}{p(q_t = k)} = \frac{p(x_t \mid q_t = k)}{p(x_t)}, \qquad (2)$$

These scaled probabilities are trained using discriminate properties of ANN. During the decoding the resulting scaled probability divisor $\frac{p(x_t \mid q_t = k)}{p(x_t)}$ does not depend on class and can be used as normalized constant. Thus scaled probability can be used in Viterbi algorithm for global *scaled probability* calculation:

$$\frac{p(X \mid M, \Theta)}{p(X)} = \sum_{paths} \prod_{t=1}^{T} \frac{p(x_t \mid q_t)}{p(x_t)} p(q_t \mid q_{t-1}), \qquad (3)$$

where the summation performs for by all ways T of model M.

These hybrid HMM/ANN methods provide more discriminant estimations for HMM emissive probabilities without applying strict hypothesizes about statistic distributions of input data.

3. EXPERIMENT

3.1. RUSSIAN SPEAKERS DATA BASE

The special speech data base was collected and prepared for experiment, this data base includes 520 phonetic balanced Russian phrases. The context-dependent phoneme – allophone was selected to be the minimum acoustical unit to describe Russian speech. Full allophone alphabet describes whole speech variety. Therefore such data base guarantees that it includes all phoneme types and can be used for valid estimation of segmentation algorithms.

Data base includes 53 speaker, 18 female and 35 male. All phrases were previously handled by speech detector and contain clear speech. Record start point corresponds to the beginning of. For each record its phonetic transcription is known.

Data base can be used in following tasks:

- Training hybrid HMM/ANN models in uncontrolled mode for speech phonemes segmentation task
- Testing of speech phoneme segmentation system on real data
- Training and testing sex identifier segmentation system

3.2. HYBRID PHONEMES RECOGNITION SYSTEM

Training of Russian speech segmentation system was done on acoustically balanced Russian speech data base. In order to model local posterior probabilities using ANN it is necessary to use as much training data as possible. This data has to be uniform that is for each phoneme it is necessary to have a lot of features vectors. The collected data base of Russian speakers fulfills all these conditions.

First stage of any speech processing system is acoustic parameterization. Most of hybrid recognition system use cepstral analysis. Experiments have been shown that cepstrum is the most effective speech parameterization feature for such task. Therefore cepstral analysis was used in this paper. In Fig 1 the typical cepstrum type is shown. In Table 1 it is detailed information about used cepstral analysis:

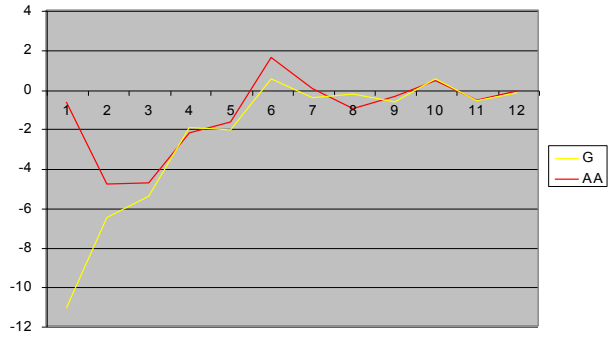

Fig.1 – Cepstrum of phonemes G and AA.

Table 1. Details of cepstrum analysis

Parameter	Font size and style
Window length	30 ms
Shifting	3 ms
Feature vector	12 cepstral coefficients
Additional parameters	Derivative 1st and 2nd power
Scope	4 adjacent wondows
Total number of components	195

After speech is represented as a sequence of observation vectors (cepstrum vectors), it is necessary to estimate local posterior probabilities using ANN. ANN outputs corresponds to phonemes. The phonetic alphabet of experimental system is presented in Table 2:

Table 2. Phonetic alphabet

U	I	O	A	E	Y	T	T’	D
D’	S	S’	Z	Z’	C	C’	SH	SH’
ZH	ZH’	N	N’	P	P’	B	B’	F
F’	V	V’	M	M’	K	K’	G	G’
X	X’	R	R’	L	L’	CH	J	J’

This alphabet is used in speech synthesis systems and successfully well shown phonetic structure of Russian language. Three-layered perceptron was used as ANN for phonetic probability estimation. Experiments shown that the size of hidden layer had to be around dimension of acoustic vector at net input. The detailed information about ANN structure and training algorithm parameters are shown in Table 3:

Table 3. ANN structure and parameters

Parameter	Value
ANN type	Three-layered perceptron with full-connecting neuron
Size of hidden layer	256
Activation for hidden layer	*HipTan*
Activation for output layer	*SoftMax*
Error	Mutal entropy
Training mode	On-line mode with forward-spot criterion

Fig 2 shows three-layered perceptron training process.

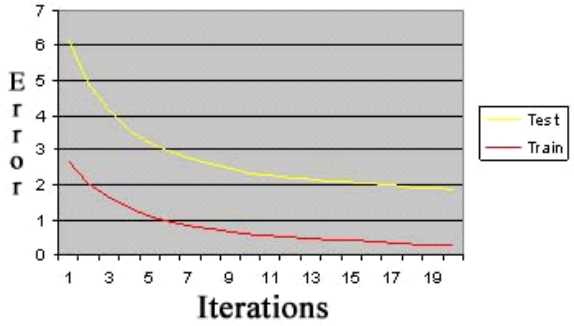

Fig.2 – Training error versus number of iterations.

In Fig 3 ANN processing results are shown. As we can see from Fig 3, well-trained perceptron can be used as segmentation system even if there is not hybrid HMM.

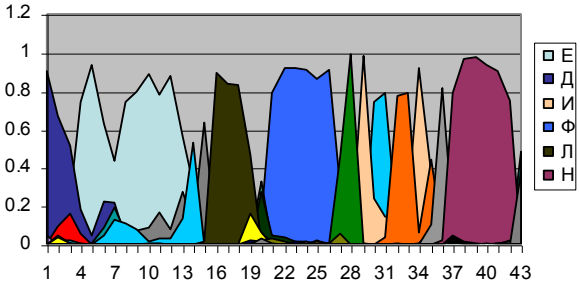

Fig.3 – ANN processing result. The word is "дельфин".

HMM topology is defined by phonetic transcription of the current speech utterance to be segmented. Common Viterbi algorithm is used to define the time of transition between states. Thus we compute rough estimation of phoneme segments border.

Testing of rough segmentation unit was done at allophone bases. Tested phrases were automatically created using speech synthesis systems. In this case phonemes borders are prior known. In Fig 4 it is shown a distribution of deviation between automatic border and handmade one. It is necessary to do explanation about this feature: X-direction is interval near known border, in which automatic border has to get, add Y-direction is percent of borders, which did not get to this interval, i.e. segmentation error.

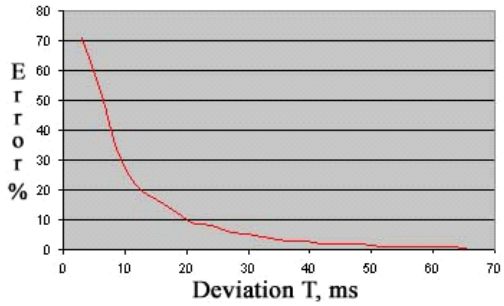

Fig.4 – Rough border deviation.

Total experimental statistic to estimate accuracy of rough segmentation unit is shown in table 4:

Table 4. Segmentation statistic using hybrid HMM

A number of segments	Average error, ms	Maximum deviation, ms
1796	9	66

4. CONCLUSION

In this paper we offer a method of speech segmentation to phonemes based on hybrid system joined Hidden Markov Model (HMM) and Artificial Neural Network (ANN). This hybrid system helps to avoid HMM disadvantages. This system has higher accuracy to define segments borders – average error is about 9 ms. The segmentation result can be improved if we use more training data for training.

5. REFERENCES

[1] Bourlard H., Morgan N. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Publishers, 1993.

[2] Deller J., Proakis J., Hansen, J. *Discrete-Time Processing of Speech Signals*. MacMillan, 1993.

[3] Gold B., Morgan N. *Speech and Audio Signal Processing*. Wiley, 2000.

[4] Jelinek F. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[5] П.Д. Кухарчик., И.Э. Хейдоров. Сегментация электрокардиограмм с помощью гибридных скрытых Марковских и моделей, *Электроника* № 3 (2008), pp. 59–62.

[6] Richard M., Lippmann R. Neural network classifiers estimate Bayesian a posteriori probabilities, *Neural Computation* №3 (1991), pp. 461–483.