PHONEME SEGMENTATION SYSTEM BASED ON HYBRID HIDDEN MARKOV MODELS AND WAVELET TRANSFORMS

Andrei Tkachenia, Alexander Soroka

Belarusian State University, Minsk, Belarus

Abstract. *In this paper we develop automatic speech segmentation system using hybrid system based on Hidden Markov Model (HMM), Artificial Neural Network (ANN) and Wavelet transforms. It was shown that the usage of ANN in order to estimate local probability in HMM leads to optimal global probability estimation in the general case. Wavelet-based feature vector provides good time resolution ability for accurate estimation of phoneme boundaries. The result of automatic segmentation is close to the manual one, and can be successfully used in real applications for speech data segmentation, training speech recognition system and text-to-speech voice data creation.*

Keywords: *Speech segmentation, hidden Markov model, artificial neural network, wavelet transform.*

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 important 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. Thirdly, accurate automatic phoneme boundaries estimation algorithms are necessary for voice data based creation for speech synthesis systems [1].

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 [2, 3, 4, 5]. 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. For overcoming some of these drawbacks, a lot of scientists integrate ANN into HMM. Neural network provides good probability for phoneme possibilities estimations for HMM states. In order to improve phoneme boundaries resolution in time it was proposed to use wavelet-based feature vector for speech signal description. Due to its unique properties wavelet transforms have found a wide application area for different scientific and technical tasks [6]. For speech analysis wavelets provides the better accuracy in transition points detection between states.

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.

HMM is based on strict probabilistic formalism, which is made them difficult to integrate with parts of heterogeneous systems. But in paper [2, 7] it was shown that if every output ANN element is associated with state κ of state sequence $в = \{1,2,..., K\}$, it is possible to train ANN for generating a good estimated value of posterior probability of output classes. In other words, if $g_k(x_t \mid 0)$ is the output function κ of ANN, and x_t - observation vector, then $g_k(x_t \mid 0^*) \ll p(q_t = \kappa \mid x_t)$, where 0^* - 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 ofthis formalism it is necessary to review basis of a stochastic finite state machines. It can be shown that $p(M \mid X, \copyright)$ can be presented in terms of transition conditional probability $p(q_t \mid x_t, q_{t-1})$ and optimal ANN parameters © can be trained according to maXimum posterior probability (MPP):

$$p(q_t \mid x_t, q_{t-1})\copyright^* = \underset{\copyright}{\mathrm{argmax}}\, p(M \mid X, \copyright) = \underset{\copyright}{\mathrm{argmax}}\, p(X \mid M, \copyright)\, p(M \mid \copyright), \qquad (1)$$

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.:

$$\frac{p^{(}qt =^{k} \backslash {}^{x}t>}{p^{(}qt =^{k)}} = \frac{p^{(x}t^{|}\ qt =^{k)}}{p^{(}V} \quad ' \qquad (2)$$

These scaled probabilities are trained using discriminate properties of ANN. During the decoding the resulting scaled probability divisor $\frac{p(xt\ 1\ qt\quad)}{p^{(}\ V}$ 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:

$$\underset{p(X)}{ZffiM®)} = \underset{paths}{1}\quad \underset{t=1}{0}\quad \underset{p(x\ t)}{{}^{Л}}p(qt\backslash qM), \qquad (3)$$

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

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

3. IMPROVED ALGORITHM FOR WAVELET TRANSFORM

The theory of wavelet transform was arranged in 90th, is not less general field of use, than common Fourier transform. Basic principle of orthogonal splitting by solid waves is based on possibility of its independent function analysis into several scaling. Wavelet-representation of signals (time function) is middle between full-spectral and -time representations.

Unique properties of wavelet-based transforms provide wide possibilities in accurate speech parameters estimation.

The continuous wavelet transform (CWT) of *f*(t) can be presented as:

$$\overset{+ж}{\underset{—ж}{Wf(u,\ \mathrm{s})=\ jf(\ ()\ us\ (t)dt,}} \qquad (4)$$

where wavelet ц _ function with zero mean and stretch parameter s and shift parameter *u* :

$$Wu,\ s \quad \frac{1}{\mathrm{Vs}}\overset{=-/=}{}Ц \quad \frac{(\mathrm{t\_u}"}{s} \qquad (5)$$

In our work we have used for CWT calculation algorithm, which implement Morlet wavelet as time-frequency functions. Firstly, we used binary version of this algorithm based on powers of 2, to achieve the highest rate. The scale parameter *s* was changed as $s = 2^{a}2^{j'J}$, where *a* - current octave, *J* - number of voices in an octave. We used *J* = 8. Secondly, the pseudo-wavelet was realized, which combines the averaging power of Fourier transform and accuracy of classical wavelet transform. We used exponential change of base frequency and linear change of window size. This leads to the full correspondence of frequency scales of wavelet and pseudo-wavelet transform. In this case (4) transforms to:

$$\overset{+\ ж}{\underset{\_ж}{{}^{W}pseudo^{f}(\mathrm{X}^{s)} =\ \mathrm{j}^{f(t)}Ps^{((\_\ u\ )\ d\ t}\ \bullet}} \qquad (6)$$

where *ps (t)* is a complex pseudo-wavelet with base frequency coordinated with wavelet frequency in scale *s* .

The usage of pseudo-wavelets lets to average non-informative signal deviations during feature vector forming. In such a way we achieve higher accuracy for high frequency analysis then it can achieved using FFT.

4. EXPERIMENT

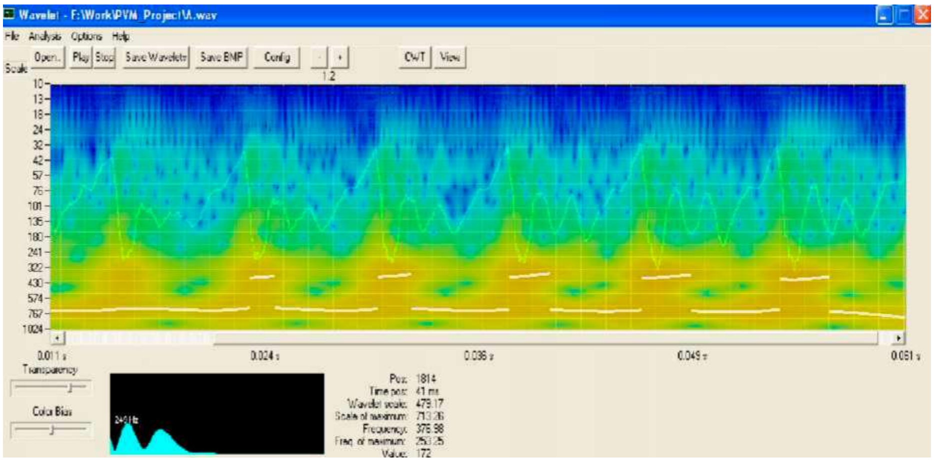
4.1. Phoneme Data Base

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 speakers, 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.

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. The mel-cepstrum features were created for each speech segment based on wavelet transform. Typical wavelet image for phonemes is shown at Fig. 1.



Pos: 1814
Time pos: 41 ms
Wavelet scale: 479.17
Scale of maximum: 713.26
Frequency: 376.98
Freq. of maximum: 253.25
Value: 172

Д____ЦУЗ

Fig. 1. Wavelet transform of phoneme «A»

Following alphabet (Table 1) is used in speech synthesis systems and successfully well shown phonetic structure of Russian language:

Table 1

Russian 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'

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 2:

Table 2

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

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 created and manually segmented to phonemes by specialists. In this case phonemes borders are prior known. So it was found a distribution of deviation between automatic border and handmade one. Total experimental statistic to estimate accuracy of rough segmentation unit is shown in Table 3.

Table 3

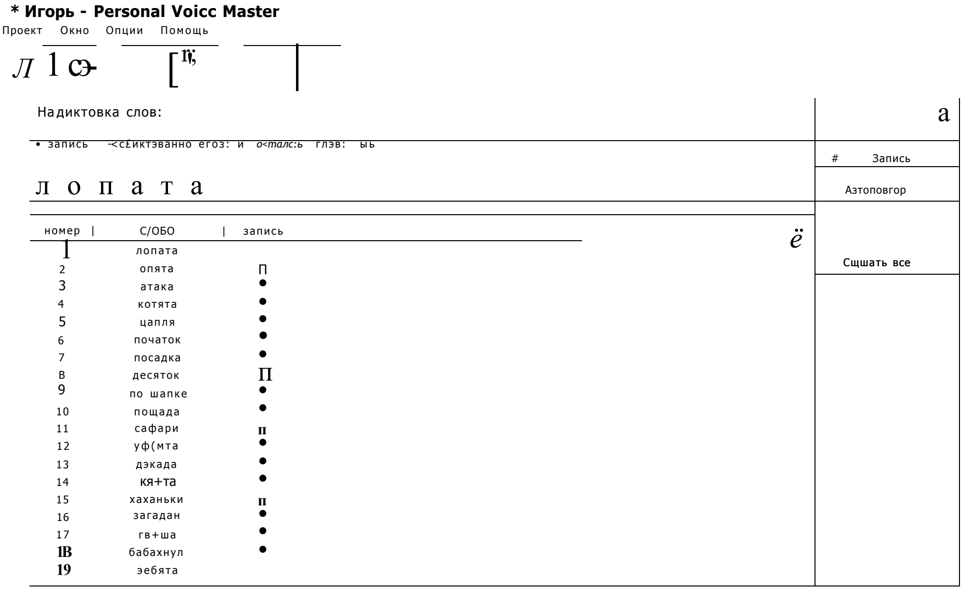
Segmentation statistic using hybrid HMM

Technique	A number of segments	Average error, ms	Maximum deviation, ms
HMM/ANN	1796	9	66
HMM/ANN+ pitch correction	1796	3	45

Depending on the type of phoneme (voiced, unvoiced, sonant, etc.) the segmentation resolution of the system varies from 2 up to 66 ms. The rough boundaries were processed by additional correcting technique that let to put the boundary mark to the pitch start point.

4.2. Software Application «Personal Voice Master»

Described theory was successfully used in software application "Personal Voice Master". This software was developed for creation ofthe acoustic data base compatible with Sakrament TTS engine in order to provide the possibility of automatic custom TTS voice design. In this program phoneme segmentation algorithm described earlier is used for creating and selection of best phonemes and saving their pronunciations in a data base. By the first stage the program records proposed to us a set ofwords, which we spoke into microphone Fig. 2. After that it is made phonemes transcription ofthese words and their automatic phonemes segmentation. In the issue we have a set of phonemes with their various pronunciations.


* Игорь - Personal Voicc Master
Проект Окно Опции Помощь
л 1 ↻
[ⁿ;
Надиктовка слов:
a
• запись -<c£иктэванно егоз: и о<талс:ь глэв: ыь
# Запись
Азтоповгор
ё
Сщшать все

л о п а т а

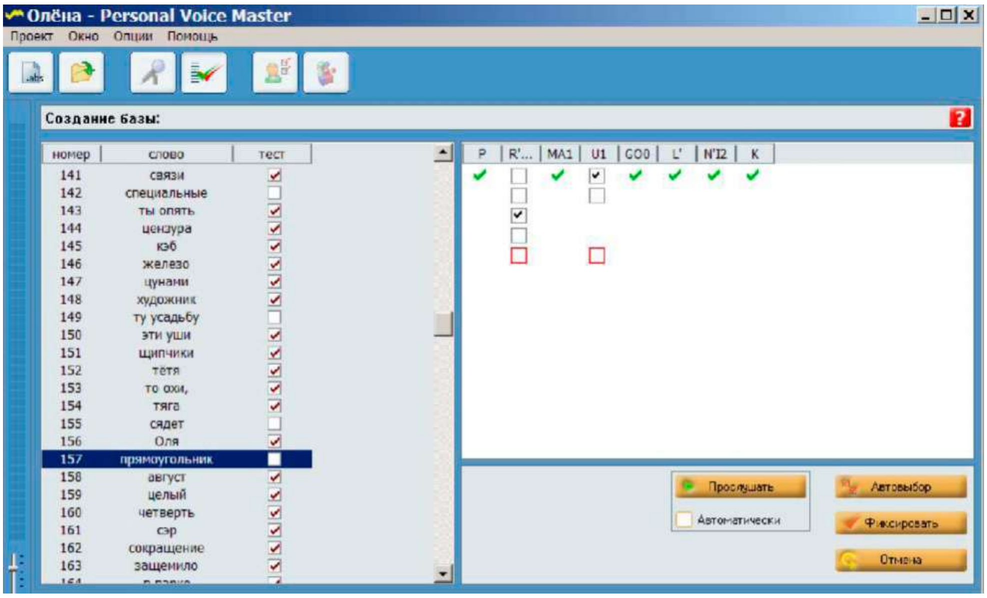
номер	С/ОБО	запись
I	лопата	
2	опята	П
3	атака	●
4	котята	●
5	цапля	●
6	початок	●
7	посадка	●
В	десяток	П
9	по шапке	●
10	пощада	●
11	сафари	п
12	уф(мта	●
13	дэкада	●
14	кя+та	●
15	хаханьки	п
16	загадан	●
17	гв+ша	●
IB	бабахнул	●
19	эебята	

Fig. 2. Words recording in software application «Personal Voice Master»

By the next stage the program proposes a test words set, for which it has already done phonemes transcription. By this stage we have all variant of phonemes pronunciations, which were got at the previous stage. User can choose the best one pronunciation from the proposed (vertical grey squares), beforehand he listen the result. After that it is necessary to settle the chosen phoneme pronunciation (green tick under phoneme transcription) Fig. 3. Now phonemes alphabet oflanguage can be created.

At the final stage, user can print any word, the program automatically make word phonemes transcription and each phoneme is given their proper pronunciation from the phoneme alphabet. Thus we have voice synthesizer tune in to a particular speaker.

Using the presented software there were done 15 voice data bases for different speakers with good synthesis quality, this fact additionally proves that hybrid HMMs with wavelet-based features vectors are very successful technique for phoneme segmentation task.



Создание базы:

номер	слово	тест
141	связи	☑
142	специальные	□
143	ты опять	☑
144	цензура	☑
145	кэб	☑
146	железо	☑
147	цунами	☑
148	художник	☑
149	ту усадьбу	□
150	эти уши	☑
151	щипчики	☑
152	тетя	☑
153	то охи,	☑
154	тяга	☑
155	сядет	□
156	Оля	☑
157	прямоугольник	□
158	август	☑
159	целый	☑
160	четверть	☑
161	сэр	☑
162	сокращение	☑
163	защемило	☑

P	R'...	MA1	U1	GO0	L'	N'I2	K
✓	□	✓	☑	✓	✓	✓	✓
	□		□				
	☑						
	□						
	□		□				

Fig. 3. Voice creating in software application «Personal Voice Master»

5. CONCLUSION

In this paper we offer a method of speech segmentation to phonemes based on hybrid HMM/ANN system with wavelet-based feature vectors. This hybrid system helps to avoid HMM disadvantages and provide good time resolution for precise phoneme boundaries estimation. This system has higher accuracy to define segments borders - average error is about 9 ms, the mean resolution ability improved up to 3 ms if the pitch start point correction procedure is used. The proposed method was applied in realized in software "Personal Voice Master", the software usage experience additionally proves that hybrid HMMs with wavelet-based features vectors are very successful technique for phoneme segmentation task.

References

1. *Long, Q.* / Qin Long, Yi-Jian Wu, Zhen-Hua Ling, Ren-Hua Wang, Li-Rong Dai // Proc. ofIEEE International Conference on Volume. 2008. P. 4621-4624.

2. *Bourlard H.* Connectionist Speech Recognition - A Hybrid Approach / H. Bourlard, N. Morgan. Kluwer Academic Publishers. 1993.

3. *Deller, J.* Discrete-Time Processing of Speech Signals. MacMillan. 1993.

4. *Gold, B.* Speech and Audio Signal Processing. Wiley. 2000.

5. *Jelinek, F.* Statistical Methods for Speech Recognition. MIT Press. 1998.

6. *Siafarikas M.* Speech Recognition using Wavelet Packet Features // Journal of Wavelet Theory and Applications. 2008. Vol. 2. No 1. P. 41-59.

7. *Richard, M.* Neural network classifiers estimate Bayesian a posteriori probabilities / M. Richard, R. Lippmann // Neural Computation. 1991. № 3. P. 461-483.