International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 7, July - 2013

Two Step Algorithm Of Separate Words Recognition In Continuous Speech Based On Usage Of Acoustic And Language Models

Igor Cheydorov
Belarussian State University, department of Radiophysics and Digital Media Technologies

Aliaksei Kuzmin
Belarussian State University, department of Informatics and Computer Systems

Abstract

In this article the new separate words search and recognition approach is proposed. It is based on successive application of acoustic models that allows evaluate the probability of the corresponding phoneme observation along the signal. The received noisy phoneme sequence is used to spot the most probable recognised word. To tackle this issue the Hidden Markov models (HMMs) formalism is employed. The comparison with the baseline algorithm "Token Passing" has showed the simplicity and effectiveness of the proposed method for the system with a restricted dictionary.

1. Introduction

The role of the decoder in a speech recognition system is to find the optimal word sequence $\hat{W}$ given the sequence of acoustic feature vectors X using the information from the acoustic model $P_A(X|W)$ and lanquage model $P_L(W)$ via the Bayes decision rule:

$$\hat{W} = \text{argmax}_w P_A(X|W) P_L(W) \qquad (1)$$

The most successful decoding algorithm uses weighted finite-state transducers (WFSTs), which allow to efficiently encode a plenty of prior information (acoustic model, language model, and HMM topology). The network composed from WFSTs, after optimisation, is directly used in a time-synchronous Viterbi decoder [1]. Such algorithm significantly outperformed the classical approaches [2].

The composition of WFSTs is implemented in static graph created by successively expanding words in the n–gram model by the corresponding transcriptions according to the different pronunciation variants.

The main advantage of usage of the static graph is that it can be compiled and optimized only once before the recognition [1], so only minimal time resources are consumed during decoding.

However, creation and optimization of the graph becomes computationally challenging when dictionary is large. For instance, it requires approximately 35 h to compile the search network for the Arabic speech recognition system with the vocabulary of about 2.5 million words [3].

Beside it, every phoneme acoustic model is rigidly connected within graph, in which case it becomes challenging to use prior information about presence or absence of definite class phonemes. For example, it is impassable to encode the information that some signal portion contains only vowels or only consonants.

In this article the alternative decoding algorithm is proposed. It's main principle is to separate in time the application of acoustic and language models. On the first step acoustic models are successively used to evaluate the log probability of the corresponding phoneme observation along the whole signal.

It is followed by the second step, on which language model is utilized to look up the word that suits the most to the noisy transcription received from the first step.

What presents the language models are HMMs which are trained on different pronunciation variants of the corresponding word. The training sequence includes transcriptions with the most widespread mistakes (deletion, substitution, and input mistakes) in recognition of the reference word. Such approach enhances the flexibility of the recognition system and substitutes tight search network by HMMs for every word in vocabulary.

2. The evaluation of phoneme observation probability along signal

As mentioned above the goal of the first step of the decoding is to find the most appropriate phoneme sequence given the signal. Here acoustic model is employed to evaluate the log probability of the phoneme observation for every successive time frame (fig. 1).

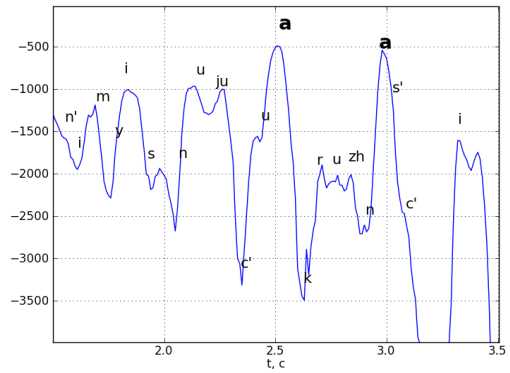

Figure 1. The log probability of the phoneme "a" observation along the part of the signal

Plots of the most probable particular phoneme observation are found using previously calculated threshold $P_{th_{ph}}$. The condition for that is the following:

$$P\left(X\left(t,T_{ph}\right)\middle|\Lambda_{ph}\right) > P_{th_{ph}},$$
$$X\left(t,T_{ph}\right) = x_t, x_{t+1}, \dots x_{t+T_{ph}} \qquad (2)$$

where $X\left(t,T_{ph}\right)$ denotes the acoustic feature vectors sequence that is started from moment t, T_{ph} is the particular phoneme sequence length, $P\left(X\left(t,T_{ph}\right)\middle|\Lambda_{ph}\right)$ is the probability of the observation sequence $X\left(t,T_{ph}\right)$ given HMM parameters Λ_{ph}.

Phoneme is associated with the time moments that are the middles of the time axis plots where log probability of this phoneme observation exceeds the corresponding threshold. After that all phonemes found in signal are ordered according to their time moments. The result is the most appropriate phoneme string, which in most cases will be noisy.

In the very beginning the probability of silence observation is evaluated in order to find the borders of the pronounced words. Other phonemes are being searching within plots defined by these borders. Such technique allows significantly reduce the amount of computation and efficiently handle silence spots. The value of lob probability threshold$P_{th_{ph}}$ as well as of observation sequence length T_{ph} is defined according to the expert assessment.

3. Selection of word which matches the most to the noisy transcription

The result of the phoneme searching within signal during the first step is a noisy phonetic representation of the pronounced utterance.

On the second step language model operates on this string in order to pick up word from the vocabulary which transcription best matched to the obtained sequence.

Essentially it is a problem of string comparison, which has been in focus of computer sciences for a long time [4, 5]. Computational biology is one of the most successful field which is equipped with this knowledge. It is exploited for the DNA chains comparison in order to reveal regions that encode the definite biological features [6].

The same principle is employed in the described work. HMMs are used to create statistical model $\Lambda_{ph} = \pi_i, a_{ij}, b_i$ [7] for every single word in the vocabulary. Every model is trained on different variants of the corresponding word pronunciation. During the decoding these models are competing on a given noisy string obtained after the first step according to (3). The recognized word $\hat{W}$ corresponds to the model $\hat{\Lambda}$ with the highest likelihood.

$$\hat{\Lambda} = argmax_{\Lambda}\left(\sum_{S=\{s_t\}}\left[\pi_{s_1} b\left(x_1\middle|\Lambda_{s_1}\right)\prod_{t=2}^{T} a_{s_{t-1}s_t}\, b\left(x_t\middle|\Lambda_{s_t}\right)\right]\right) \qquad (3)$$

here $S = \{s_t\}$ denotes HMM states, π_{s_i} are the initial state probabilities, $a_{s_{t-1}s_t}$ — the state transition probabilities and $b\left(x_t\middle|\Lambda_{s_t}\right)$ is the probability to observe symbol x_t being in state s_t of the model Λ.

An overall decoding scheme is depicted on figure 2

It is worth mentioning that pronunciation variants only can not be used to form an adequate training corpus to create the word HMM. This corpus has to be expanded with the sequences which includes most widespread substitution, deletion and input mistakes for the particular word recognition. It allows it to take into account the most frequent transcription strings that correspond to the reference word.

For instance, russian word "ammiak" has following transcription from the dictionary: "a", "m' ", "i", "a", "k", but on practice in majority of cases it is recognized with the additional sound "e" between "i" and "a".

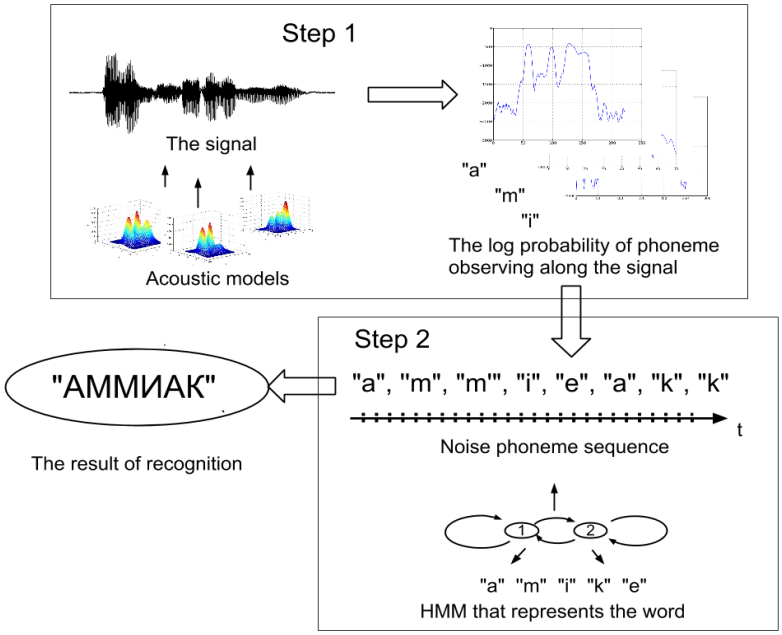

Figure 2. The scheme of two step decoding algorithm for the search and recognition of separate words in continuous speech

4. Summary of the Experiment

The main goal of the experiment was to check up is it possible to reach the same recognition accuracy comparing with baseline system that operates on the search network [8]. Under this condition the computational consumption of the new decoder in terms of long operations is required to be comparable with the baseline system.

The training corpus consists of one speaker records with total running time about an hour. Fifty nine russian words form the testing vocabulary.

Acoustic signal is divided into frames by successive application of 25 msec Hamming window with 10 msec time shift. Every frame is transformed into mel-frequency cepstral coefficients (MFCC).

Acoustic model for every monophone is 3 states HMM with state-dependent GMMs. Every GMM consists of 5 components with full covariance matrix [7]. Such topology allows it to increase the recognition accuracy at the phoneme level [9].

The number of frames in sequence for every model evaluation plays essential role in discriminative abilities of the first step of the decoding, but at the same time it is the two edged sword. On the one hand, the more vectors in sequence, the higher gap in the probability between regions of presence and absence of the particular phoneme. On the other hand, sequence expanding brings to rapid increase of computations. That why number of frames in the sequence was carefully defined according to the expert assessment in order to retain the balance between accuracy and velocity. The same approach helped to set log probability thresholds.

The test corpus includes records of the same speaker with total running time about 108 sec and vocabulary formed with 11 words.

Both the proposed decoder as well as the baseline system showed accuracy close to 100%. Number of long operations in two step decoder just slightly exceeds it's quantity in decoder that operates on static graph. It implies that the discussed decoding algorithm already possesses characteristics comparable with the algorithms based on usage of a search network.

The average time of one word HMM training is equal to 0.00184 sec. It can be easy deduced that the language model for 2.5 million words vocabulary can be created for just about 1.28 hour, which is much faster than 35 hour compilation of static graph as discussed in [3].

5. Conclusion and Future Work

The experiment has demonstrated the ability of the proposed algorithm to reach the recognition accuracy that is comparable with the most widespread counterparts that based on search network usage.

The striking point is that the discussed decoder allows significantly reduce time for the language model creation, along that it gives more flexibility from the view point of prior signal analysis. In particular it is relatively easy to define whether some signal region represents vowel or consonant sound just by comparing magnitude distribution over spectrum. Such preprocessing will allow it to decrease number of acoustic model evaluation which is very computationally challenging.

All in all, the proposed decoding algorithm might outperform the baseline system that operates in the static graph in terms of time consumption, but at the same time it must steel show the same accuracy level.

References

[1] M. Mohri, F. Perreira, and M. Riley, "Weighted finite state transducers in speech recognition", *Comput. Speech. Lang*. vol. 16. no. 1. 2002, pp. 69–88.

[2] S. Kanthak, H. Ney, M. Riley, and M. Mohri, "A comparison of two LVR search optimization techniques" in *Proc. Int. Conf. Spoken Language Processing (ICSLP)* -- 2002. -- pp. 1309–1312.

[3] H. Soltau, G. Saon, "Dynamic network decoding revisited" in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. -- 2009. -- pp. 276–281.

[4] P. A. V. Hall, G. R. Dowling, *Approximate String Comparison*, Computing Surveys. -- 1980. -- 12. – pp. 381–402.

[5] Jaro, M. A., "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*. -- 1989. -- 414–420.

[6] Churchill, G. "Stochastic models for heterogeneous dna sequences", *Bulletin of Mathematical Biology*. -- 1989. -- 51:79–94.

[7] Rabiner, L. R., "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proceedings of the IEEE 77 (2)*. -- 1989. -- pp. 257–286. doi:10.1109/5.18626.

[8] S. J. Young, N. H. Russell, J. H. S. Thornton, "Token passing: a simple conceptual model for connected speech recognition system", *Tech. Report*, Cambridge University Engineering Depart, 1989.

[9] Vertanen, K. Baselint "WSJ Acoustic Models for HTK and Sphinx: Training Recipes and Recognition Experiments", *Tech. Report*, University of Cambridge, Cavendish Laboratory, 2006