*Z. YESSENBAYEV*

# SOME PROBLEMS OF SPEECH RECOGNITION

In this work, we state some basic problems (robust representation of speech, denoising, acoustic database etc) which need thorough consideration before any speech recognition algorithms are designed. After crafting these problems we can build and test a speech recognition systems. Although the problems are stated in general form and based on the foreign experiense, no doubts, that they are applicabale to Kazakh language, which we aim to have as an object of our future research

## Introduction

Speech is an interesting and attracting object of study for such areas as linguistics, neurobiology, physics as well as electrical engineering and computer science. This is because speech is a multi-facet object and yet not well understood. One of the important aspects of speech for computer science is speech recognition, which is an attempt to automate the "understanding" of speech by machines. The ability of a computer to "understand" speech and act accordingly would potentially reduce the human-load and the risk in human-dependent applications. The real applications of automatic speech recognition systems (ASR) are widely used in telecommunication, medicine, and military in many of the developed countries such as USA, Russia, Japan, etc. To contrast, there are almost no such real systems designed for Kazakh language.

Thus, the object of the current study is Kazakh speech and its phonetic diversity. The leading universities of our country, including Eurasian National University, intensively conduct the researches in this area for the last decade. However, there are no common standards regarding the phonetics of Kazakh language, there are no public databases, which would include the utterances of different speakers (adult/children, male/female) of different dialects, and as a result, there is no opportunity to adequately compare the results among the academics. Definitely, all these decelerate the process of the knowledge accumulation in this area. Therefore, we pursue the initial goal of more detailed analysis of the phonetics of Kazakh language in terms of speech signal, the extraction of the spectral characteristics of speech for the different dialects, the gathering of the statistical data, on the basis of which further research would be made. Only after the clear understanding of the object and having gathered enough data we consider the development of the methods and algorithms for automatic recognition of continuous Kazakh speech. But before all we state some fundamental problems to cope with based on the experience of the leading countries in this area. Clearly, without lost of generality, most of the ideas are applicable regarding Kazakh language.

## A data representation problem

Although, the success of the real ASR systems is notable (up to 95%), they all suffer from noisy environment and the performance degrades significantly. And because of this instability, they hardly can substitute human in the applications where the responsibility is high. For this reason, the robustness of such systems plays very important role. Interestingly, in the same noisy conditions human can still understand speech, for example, the ability of human to focus on one of two speakers and hear only his/her speech and regard the other speaker as noise. This is what the modern speech recognition systems can't do. One question that consequently arises from this example is that is there a unique representation of speech signal which helps human extract only the information he needs? So the basic concern is the quest for better (in the sense of robustness) description of speech, which will benefit the current systems and prevent them from failing in the presence of noise, that is, will provide some stability and reliability in the recognition process.

There are different approaches to process a speech signal, and two major of them are DTW-based and HMM-based approaches. The dynamic time warping (DTW) method is based on finding the similarities between two time sequences that change in time and speed, for example, matching two sequences with different speaking speeds [7]. The basic algorithm that this method exploits is a dynamic programming, hence the name. In spite of its algorithmic power, DTW was mostly replaced by the second approach, which proved to be far more successful. The HMM-based methods build statistical model of a continuous time-varying speech signal, viewed as a sequence of smaller slowly time-varying fragments, called frames, thus approximating the signal as a stationary process having Markov property [4, 9]. It is also worth mentioning traditional machine-learning algorithms such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) as well as landmark-based approaches [1, 2]. The important aspect of these methods is the representation of signal within each frame. Such representations are called feature sets. The most common feature sets are cepstral and perceptual linear predictive coefficients (PLP). Although they adequately describe the signal what makes these methods successful, it turns out they are not stable if the signal is noisy (~10dB SNR). Another representation of a signal can be in terms of acoustic parameters (AP) such as periodicity, spectral energy, Weiner entropy etc.

There are various trends in speech recognition community differing in algorithms and feature sets used. The approach proposed by A. Jansen and P. Niyogi [2] uses the notion of distinctive features and exploits the idea of landmark detectors, which, they claim, could be an alternative to the modern HMM-based approaches. A hierarchical model is based on the number of feature detectors that output a set of candidate landmarks, which are, then, probabilistically integrated to construct the most likely sequence of broad classes. Although the model built is up to the broad-class level (vowels, consonants, nasals, fricatives etc [6]), it explicitly includes the sonorant-obstruent segmentation stage. For the segmentation, an SVM was trained on 39-dimensional mel-frequency cepstral coefficient (MFCC) feature sets. The accuracy was estimated with the measure they suggest, which is, basically, the percentage of the phonemes that fall into corresponding sonority regions given the threshold of being "accepted" by those regions (for details, see the next section). In spite of the high performance for different thresholds, what interested us most is that the difference between the respective sonorant and obstruent measures is high and grows significantly as the threshold is increased. One of the reasons could be that the segmenter assigns wider regions for one class while narrower for the other, what makes such a difference in the measures. Therefore, the question of the quality of the segmentation arises. Another issue that was not considered in this paper is noise.

Another work [3] also builds a frame-based SVM classifier using the general-purpose MFCCs, however, the problem of noisy condition is addressed. For that, authors estimate signal-to-noise ratio from the frame energy histograms, noticing that, for stationary noise, there is going to be two peaks: one corresponds to the accumulations of non-speech frames containing only noise, and the other – to the speech plus noise portion. The difference between these two peaks gives a measure that is a "good indicator" of SNR. Then, based on this measure they vary adaptively the parameter λ in the classification rule $x_i \in \{sonorants\} \Leftrightarrow w_0^T x_i + b_0 > \lambda$ . Unclear part of this work is a map from SNR to optimal threshold during the training stage. Another problem is that the comparison of the results for noisy data was made between different settings of λ, but not with the results obtained on clean data. For example, the difference in performance between clean and pink noise added data goes above 10%, what requires the explanation how "good" or "bad" it is, i.e. the measure of accuracy estimation was not given. In addition, this type of approach doesn't take into account the noisiness of the extracted feature sets.

An alternative to these two papers in terms of data representations is a work by A. Juneja and C. Espy-Wilson [13]. Their method is based on the extraction of different acoustic parameters and passing the relevant parameters to SVMs trained for each broad class. Although it is not very clear how some of the parameters are obtained (third formant of a speaker, or probability of voicing), the idea of separating the parameters according to the broad class should really be appreciated. The attempt to compare the performances of their system with HMM-MFCC-based one fails in that they built not quite a competitive model for the second system, which has the performance not comparable to the state-of-the-art HMM-based systems.

Some other studies using APs as a basis of speech representation are described in [3, 15, 16]. All the parameters extracted are quite simple yet natural, however, the robustness to noise should be inspected carefully. For example, Wiener entropy may not reflect structural properties of a signal in the presence of noise as well as the periodicity estimation of noisy and non-stationary signals is an extremely difficult task, what shows the works such as [17, 18]. Therefore, it is not sufficient to use such a small number of parameters (3-4), for the systems that will be exposed to an adverse conditions and the noise level is considerably high.

There are also studies of slightly different manner. For instance, it is worth mentioning the work [19], which combines the power of statistics with the ideas of edge detection in computer vision. Another approach [12] uses no linguistic knowledge, but rather machine-learning algorithms based on clustering and dynamic programming techniques. In [20], a noise adaptive speech recognition system is built with acoustic models trained on noisy data, which is not common to the traditional approaches, where the training set is a clean speech.

The analysis of these works shows that cepstral-based approaches rightfully became popular among speech recognition community, but the need for more robust representation suggests augmenting them with additional cues such as acoustic parameters. The advantages of using these parameters are that they are less noise sensitive and the linguistic information extracted is meaningful and explicit for human what results in relatively better identifying and analyzing the recognition errors made by the ASR systems. And the disadvantage of them is that a good and direct description (such as periodicity or formants) of a signal requires more computational power than cepstral coefficients, therefore, mostly indirect (such as energy ratios or entropy) parameters are used, what results in lower performance even in the clean environment. So, a careful decision should be made on what type feature set to design and use – the cepstral-based coefficients or the acoustic parameters. If acoustic parameters are preferred then how to deal with the computational issues and their robustness to noise?

### A denoising problem

Another problem which needs thorough consideration and understanding is the influence of different noises (additive, convolutional) on the speech signal and the ability of their exclusion from the speech signal, i.e. signal denoising. The observations [21] show that noise can significantly change the distributions of selected parameters, for instance, the distribtion can shift or change its shape depending on the noise level and its type, what makes most of the statistical methods and algorithms almost useless. This, in turn, suggests the development of the adaptive algorithms of recognition such as [20], which would take into account the dynamics of the signal in the presence of noise, or/and the organization of the effective preprocessing of the speech signal, which would cut off the noises on the early stages. The possible techniques are the smoothing of the signal or the suppression of the noise, for which novel wavelet-based [8] or standard filtering algorithms can be applied. There are also more sophisticated algorithms such as Phase Opponency Model [23], which is an approach to a speech separation challenge (two-talker speech, speech-shaped noise). During the preprocessing it is crucial not to loose the necessary information, which would be used in later stages, while reducing as much as possible the unnecessary information in the signal.

### A database problem

Finally, to perform complete and significant research it is necessary to have qualitative and representative data. Particularly, a phonetically rich database of speech utterances is needed, which would contain the speakers of different age groups, genders, dialects in various noise conditions. Some examples of such databases for English language are TIMIT – unnoisy clear data with different speakers, NTIMIT – the same database, but transmitted through the phone lines, Aurora-2 – multi-conditional database, containing the noises of metro, cars, exhibition halls etc. Although there are no *full* analogues for Kazakh language, there are efforts to build simple test databases of the isolated words and digits [22]. The main issues when gathering the data are the quality of the utterances as well as their transcription. To achieve good results, it requres the participation of the qualified specialists in lingustics, who would prepare representative sentences to be uttered and extract their transcriptions from the speech signals. No doubts, that such a database would be useful practically and teoretically for the further researches on speech recognition and linguistics.

### REFERENCES

1. *K. Schutte, J. Glass*, Robust detection of sonorant landmarks. Interspeech, 2005

2. *A. Jansen, P. Niyogi*, A probabilistic speech recognition framework based on the temporal dynamics of distinctive feature landmark detectors. July 4, 2007

3. *P. Niyogi, C. Burges, and P. Ramesh*, Distinctive feature detection using support vector machines. ICASSP, 1998

4. *W. H. Abdulla, N. K. Kabasov,* The concept of Hidden Markov Models in speech recognition. TR 99/09, New Zealand, 1999

5. *Dietterich, T. G.* Machine Learning for Sequential Data: A Review. In Proceedings of the Joint IAPR international Workshop on Structural, Syntactic, and Statistical Pattern Recognition (August 06 - 09, 2002). Lecture Notes In Computer Science, vol. 2396. Springer-Verlag, London, 15-30. 2002

6. *J. P. Olive, A. Greenwood, J. Coleman,* Acoustics of American English speech. A dynamic approach. Springer, 1993

7. *C. S. Myers and L. R. Rabiner.* A comparative study of several dynamic time-warping algorithms for connected word recognition. The Bell System Technical Journal, 60(7):1389-1409, September 1981.

8.*M. S. Crouse, R. D. Nowak, R. G. Baraniuk, W*avelet-based statistical signal processing using Hidden Markov Models. IEEE transactions on Signal Processing, Vol. 46, No. 4, April 1998

9. *Lawrence R. Rabiner, A* tutorial on Hidden Markov Models and selected applications in speech recognition, Proceedings of the IEEE, 77 (2), p. 257–286, February 1989.

10. *Jeff A. Bilmes,* A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian Mixture and Hidden Markov Models, TR-97-021, Berkeley, CA, April 1998

11. *L. R. Rabiner, R. W. Schafer.* Englewood Cliffs, Digital processing of speech signals. London: Prentice-Hall, 1978

12. *M. Sharma, R. Mammone,* "Blind" speech segmentation: automatic segmentation of speech without linguistic knowledge. In ICSLP-1996, 1237-1240, 1996

13. *A. Juneja, C. Espy-Wilson,* Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning. In the proceedings of 9th International Conference on Neural Information Processing, Singapore, 2002, Volume 2, Page 726-730, 2002

14. *T. Pruthi, C. Espy-Wilson,* Acoustic parameters for automatic detection of nasal manner. Speech Communication, Volume 43, Issue 3, August 2004, Pages 225-239

15. *Zhimin Xie, P. Niyogi,* Robust acoustic-based syllable detection. Proceedings of Interspeech, 2006.

16. *P. Niyogi, E. Petajan, J. Zhong,* Feature based representation for audio-video speech recognition. Proceedings of the Audio Visual Speech Conference, Santa Cruz, CA, 1999.

17. *S. Parthasathy, S. Mehta, S, Srinivasan,* Robust periodicity detection algorithms. Proceedings of the 15th ACM international conference on Information and knowledge management, 2006

18. *M. Vlachos, Philip Yu, V. Castelli,* On periodicity detection and structural periodic similiarity. In Proceedings of SIAM International Conf. on Data Mining (SDM), Newport Beach, CA, 2005

19. *Y. Amit, A. Koloydenko, and P. Niyogi.* Robust acoustic object detection. J. Acoust. Soc. Am, 118(4), 2005.

20. *Kaisheng Yao, K.K. Paliwal, S. Nakomura,* Noise adaptive speech recognition with acoustic model trained from noisy speech evaluated on Aurora-2 database. In ICSLP-2002, 2437-2440, 2002

21. *Yessenbayev Z. A.* Robust sonorant-obstruent segmentation of speech signal. Master Thesis, The University of Chicago, 2008

22. *Saparaliyev A. N., Yessenbayev Z. A.*, The designing of the acoustic database of the digits of kazakh language and the development of the methods and algorithms for their automatic recognition. Bachelor Thesis, Eurasian National University, Astana, 2009

23. *Deshmukh O., Anzalone M., Espy-Wilson C., Carney L.,* A noise reduction strategy for speech based on phase-opponency detectors. 149th Meeting of the ASA, 2005.

### Резюме

Бұл жұмыста сөйлеу тану үшін қажет кез келген алгоритмді жасау алдында түсінуді талап ететін негізгі проблемалар (сигналды тұрақты бейнелеу, шудан құтылу, акустикалық деректер базасын жасау және басқа талқыланады. Тек айтылған проблемаларды шешкеннен кейін ғана біз сөйлеуді тану жүйесін құрастыра және тестілей аламыз.

### Резюме

В данной работе обсуждаются основные проблемы (устойчивое представление сигнала, избавление от шумов, создание акустической базы данных и др), которые требуют серьезного осмысления перед разработкой любых алгоритмов по распознаванию речи. Только после решения отмеченных проблем мы можем строить и тестировать систему распознавания речи.

UDK 004.432.4

*Евразийский национальный университета им. Л.Н. Гумилева* *Поступила 23.09.09 г.*