Journal of Physics: Conference Series

Implementation of the Intelligent Voice System for Kazakh

To cite this article: Zh Yessenbayev *et al* 2014 *J. Phys.: Conf. Ser.* **495** 012043

View the article online for updates and enhancements.

Implementation of the Intelligent Voice System for Kazakh

Zh Yessenbayev[1], N Saparkhojayev[2], T Tibeyev[3]
[1]Nazarbayev University Research and Innovation System, 53 Kabanbay Batyr Ave., Astana. 010000, Republic of Kazakhstan
[2]ISMB Research Institution, POLITO, Torino, 10129, Italy
[3]Suleyman Demirel University, 1/1 Abylaikhan St., Kaskelen, Almaty, 040900, Kazakhstan

E-mail: zhyessenbayev@nu.edu.kz

Abstract. Modern speech technologies are highly advanced and widely used in day-to-day applications. However, this is mostly concerned with the languages of well-developed countries such as English, German, Japan, Russian, etc. As for Kazakh, the situation is less prominent and research in this field is only starting to evolve. In this research and application-oriented project, we introduce an intelligent voice system for the fast deployment of call-centers and information desks supporting Kazakh speech. The demand on such a system is obvious if the country's large size and small population is considered. The landline and cell phones become the only means of communication for the distant villages and suburbs. The system features Kazakh speech recognition and synthesis modules as well as a web-GUI for efficient dialog management. For speech recognition we use CMU Sphinx engine and for speech synthesis- MaryTTS. The web-GUI is implemented in Java enabling operators to quickly create and manage the dialogs in user-friendly graphical environment. The call routines are handled by Asterisk PBX and JBoss Application Server. The system supports such technologies and protocols as VoIP, VoiceXML, FastAGI, Java SpeechAPI and J2EE. For the speech recognition experiments we compiled and used the first Kazakh speech corpus with the utterances from 169 native speakers. The performance of the speech recognizer is 4.1% WER on isolated word recognition and 6.9% WER on clean continuous speech recognition tasks. The speech synthesis experiments include the training of male and female voices.

1. Introduction

Analysis of several speech recognition and synthesis engines such as Sphinx4, HTK, Julius, FreeTTS, Festival, etc., showed that they require solid background knowledge in speech technologies to configure and get a functional system. However, this is a difficult task for most of the end-users such as governmental, educational or other industry entities that may not have adequate specialists. Moreover, to our best knowledge, there are no deployments of speech-based services in Kazakhstan to date. All the current call-centers use DTMF-based software. In view of this, we started a project of developing a prototype of an intelligent voice system (IVoS) to help fast and easy deployment of speech-enabled contact-centers and information desks for those who lack special knowledge in speech technologies.

2. System description

The intelligent voice system features Kazakh speech recognition and synthesis modules as well as a web-GUI for efficient dialog management. The high level architecture of the system can be outlined as in Figure 1. At the core of the system, there is an application server that manages the system's data and work flows as well as access to other modules. The server consists of a Control module that handles run-time work flow and an administration web-GUI that enables operators to quickly create, manage and browse the status of the dialogs in user-friendly graphical environment. The dialogs are compatible with VoiceXML format. For speech recognition we use CMU Sphinx engine and for speech synthesis – MaryTTS. However, the system defines several own interfaces to interact with the third-party speech engines based on JavaSpeech API, so that one can prefer other solutions. The call routines are handled by Asterisk PBX and forwarded using FastAGI protocol to our remote application server. All application related data are stored in MySQL database and accessed via Java Persistence.

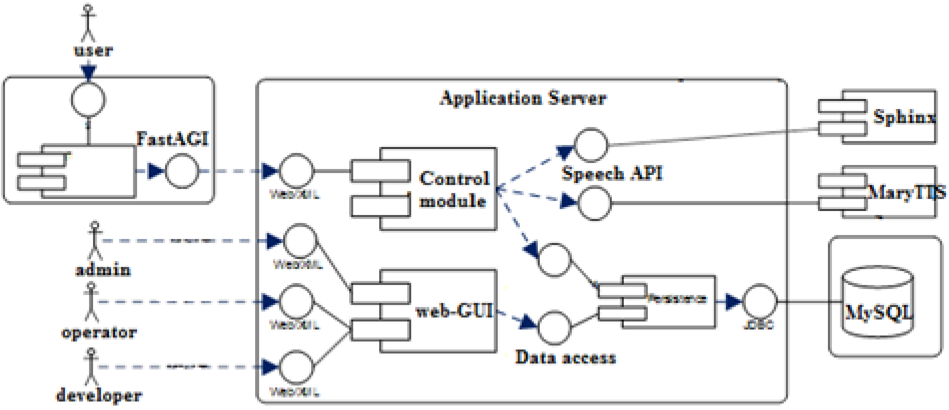

Figure 1 System architecture

3. Speech recognition and synthesis

3.1 Data

Most modern state-of-the-art speech recognition systems require audio data for training the acoustic models. Depending on the type of an application data needed varies from high quality microphone speech (WSJ0) to telephone speech (Switchboard or CALLHOME), from continuous speech (TIMIT) to connected (TIDIGITS) and isolated words (PhoneBook). In our current work, we compiled the first Kazakh speech corpus which was initiated as a part of the Kazakh Language Corpus presented in [1]. The speech corpus consists of three parts: 1) telephone speech of isolated and connected words (names, cities, dates, etc.) for the medium vocabulary (up to 2000 words) command-and-control applications; 2) high quality microphone speech of 169 native speakers for the large vocabulary continuous speech recognition; 3) long recordings of two professional actors for the speech synthesis experiments. Audio data were captured using professional vocal microphone as well as our call-collecting system using Asterisk PBX built on top of the existing university-wide infrastructure. All the recorded audio files were manually post-processed to have each utterance in a separate file. The total duration of the audio files is about 60 hours. The speakers that took part in the recordings are volunteers were balanced by region, age and gender. There are 15 region groups: 14 official regions ("oblast") of Kazakhstan and one group for those who lived outside of the country. The age groups are divided into four ranges between 18 and 65 years. We tried to keep the number of speakers of one gender per profile not more than 3. Additionally, two professional actors, one male and one female, were selected among these speakers to read excerpts from news articles and novels. The following Table 1 presents the distribution of the recorded speakers across the regions, gender and age groups.

Table 1. The distribution of the speakers

Age group	I		II		III		IV		
Region	F1	M1	F2	M2	F3	M3	F4	M4	Sum
1	3	3	2	1	2	1	2	1	15
2	2	3	2	1			2	1	11
3	1	1	2	3	2	1	1		11
4	3	2		1		1			7
5	2	2	2	1	2	2	2	1	14
6	2	2	2	2	2		1	2	13
7	2	2	1	2	2		2	1	12
8	2	1	1	2	1	1	2	1	11
9	3	2	2	1	3	1	1	1	14
10	1	1	2	2	1	1	2	1	11
11	2	1	2	1	1		2		9
12	2	2	2		2	1	2	1	12
13	2	2	2	1	1	1	1	1	11
14	2	1	1	1	1	2	1	2	11
15	1	3		1	2				7
Total	30	28	23	20	22	12	21	13	169
	34%		25%		20%		20%		

3.2. Experiments

Our first experiment was conducted for the isolated and connected words recognition. An acoustic model was trained using CMU Sphinx toolkits. In the experiments we aimed to build an adequate baseline acoustic model to be used in our intelligent voice system. Since most of the parameters in Sphinx are optimally pre-configured, there were basically two parameters that we were interested in to improve the recognition performance: LDA dimension and number of tied states (senones). Tables 2 and 3 show the results of the experiments conducted.

The front-end module was set to output default parameters such as 13 mel-frequency cepstral coefficients with their first and second derivatives. Additionally, speaker adaptation techniques such as CMN [2], LDA [3] and MLLT [4] are performed on feature vectors. We used a context-dependent tied-state continuous Hidden Markov Model with 8 Gaussian mixtures per state. The total number of test words is 5052. The dictionary is compiled from the text materials used for telephone speech and contains 1684 words with their spellings as a phonetic transcription, since the orthographic transcription of Kazakh roughly corresponds to a broad phonetic transcription.

It can be seen from the tables that the optimal size of reduced space was found to be 23 as opposed to the system's default value 29. As for the senones, the best result was obtained with 2000 senones.

In the next experiment we used our data for the large vocabulary continuous speech recognition task. Here we used almost the same parameters for the acoustic modeling as in the previous experiment. The values for the LDA dimension and the number of senones are chosen to be 23 and 2000, respectively. The audio data was separated into training and test sets. The test set is balanced based on gender and includes one representative from each region. The quantitative information about both sets is given in Table 4. For the language modelling we used our transcripts to build a trigram based model. The overall performance of recognition on test data is 6.9% WER.

Table 2. Performance for different LDA dimensions (with 2000 senones)

Dimension	WER, %
29	4,4
25	4,3
24	4,5
23	**4,1**

Table 3. Performance for different numbers of tied states

# senones	WER,%
3500	5.8
3000	4.9
2500	4.8
2200	4.4
2000	**4.1**
1500	4.8

Table 3. Distribution of data in training and test sets

	Train set	Test set
# of speakers	153	16
# of audio files	11367	1176

For the speech synthesis we used MaryTTS to train one male and one female voice. There are about 6 hours of audio data for each of the voices. The main approach is also based on Hidden Markov Models. Some NLP components such as grapheme-to-phoneme models, basic POS-tagging are built during the experiments. The work on speech synthesis is still in progress.

4. Conclusion and Future Work

In this work we implemented the intelligent voice system for the development of speech-enabled call-centres supporting Kazakh language. Also we conducted the experiments on Kazakh speech recognition and synthesis. Although the results for the speech recognition are encouraging, they are obtained on clean speech data. Therefore, our next work will concern collecting and processing the real data in office environment as well as improving the system functionality.

Acknowledgements
The work is supported by the Ministry of Education and Science of the Republic of Kazakhstan.

References

[1] Z. Yessenbayev, O. Makhambetov, and M. Karabalayeva, “Kazakh Text Corpus: Description, Tools and Statistics,” Int. scientific-theoretical conference “Modern Kazakh Linguistics: Actual Problems of Applied Linguistics” (2012), pp. 61-65.

[2] Liu, F.-h., Stern, R.M., Huang, X., Acero, R., “Efficient Cepstral Normalization for Robust Speech Recognition,” In Proceedings of the workshop on Human Language Technology (1993), pp. 69–74.

[3] Haeb-Umbach, R., Ney, H., “Linear discriminant analysis for improved large vocabulary continuous speech recognition,” IEEE Int. Conf. on Acoustics, Speech, and Signal Process (1992), vol. 1, pp. 13–16.

[4] Gopinath, R.A., "Maximum likelihood modelling with Gaussian distributions for classification," IEEE Int. Conf. on Acoustics, Speech and Signal Processing (1998), vol.2, pp. 661-664.