Development of limited-vocabulary ASR for Azerbaijani

Aydin Bagiyev
ADA University
Baku, Azerbaijan
abagiyev2018@ada.edu.az

Konul Gurbanli
ADA University
Baku, Azerbaijan
kgurbanli2018@ada.edu.az

Natavan Mammadova
ADA University
Baku, Azerbaijan
ntmammadova2018@ada.edu.az

Sariyya Nuriyeva
ADA University
Baku, Azerbaijan
snuriyeva2018@ada.edu.az

ABSTRACT

Although could be widely developed and implemented in local industries to decrease costs and increase customer satisfaction, Azerbaijan currently has not brought speech recognition systems into play. One of the potential application areas is call-based taxi ordering systems as taxi services are locally popular business where customer satisfaction is crucial for a competitive advantage. Therefore, our team develops a speech recognition system for the Azerbaijani language which could be utilized by national taxi services. The system is based on CMUSphinx, an ASR toolkit that works with Hidden Markov Model [1]. Via training 100 distinct street names of a capital city and numbers from 1 to 1000 with the help of 120 speakers, the ultimate result of 95.4% accuracy of recognition was obtained. Positive results obtained show that when the project will be put into action, call centers of local taxi companies will undergo a radical, but positive change in their business flows and productivity. The research will also spread awareness on the field of AI, and thus initiate other research projects or induce actions on the relevant fields in Azerbaijan.

KEYWORDS

speech recognition system, Azerbaijani language, CMUSphinx, call center automation, Dilmanc Imla, Google Docs Text-to-Speech

RESULTS

With phonetic dictionary, language model and acoustic model built for Azerbaijani, the ultimate result of 95.4% accuracy of word recognition was obtained via training 100 distinct street names of a capital city and numbers from 1 to 1000 with the help of 120 speakers. The accuracy of results was calculated using Word Error Rate(WER) and Sentence Error Rate(SER) metrics. WER metric is based on the number of inserted, deleted and substituted words. Let N be the number of words in the text after the recognition, then:

$$WER = \frac{S + D + I}{N}$$

where S, D and I are the number of substituted, deleted, and inserted words respectively.

SER is based on the number of errors made while recognizing phrases or sentences. Let E be the number of misrecognized phrases/sentences and N the number of all phrases/sentences, then:

$$SER = \frac{E}{N}$$

Table 1: **Results of Azerbaijani Speech Recognition System (Keenax).**

Utterances (training data)	Utterances (test data)	SER	WER	Speakers F:M
14392	972	9.3%	4.6%	42:78

To compare with similar tools, we tested Dilmanc Imla(Dictate) and Google Docs Text-to-speech tool with the same data [2, 3]. Both tools demonstrated approximately 90% accuracy while Dilmanc performed slightly better with 7% WER and 22% SER than Google's tool which resulted in 9.3% WER and 25% SER.

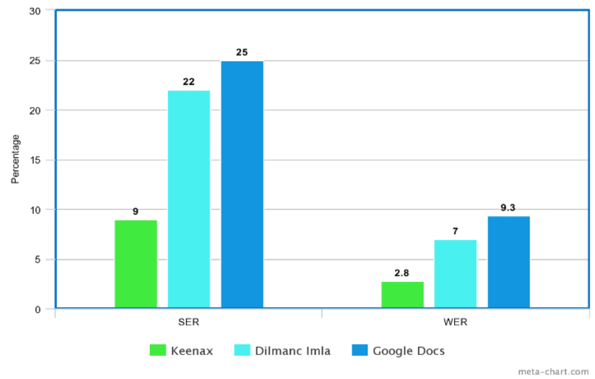

Figure 1: **Comparison of the tools.**

REFERENCES

[1] Nickolay V. Shmyrev. CMUSphinx Open Source Speech Recognition. Retrieved April 5, 2018 from https://cmusphinx.github.io/.

[2] Dilmanc Dictate. Retrieved January 2, 2018 from http://dilmanc.az/en/dilmanc-dictate/.

[3] Google Docs. Retrieved January 2, 2018 from www.google.com/docs/about/.