Piotr KOZIERSKI[1,2], Talar SADALLA[2], Szymon DRGAS[1], Adam DĄBROWSKI[1], Dariusz HORLA[2]

Poznan University of Technology, Faculty of Computing, Division of Signal Processing and Electronic Systems (1)
Poznan University of Technology, Faculty of Electrical Engineering, Institute of Control and Information Engineering (2)

doi:10.15199/48.2016.11.70

Kaldi Toolkit in Polish Whispery Speech Recognition

***Abstract**. In this paper, the automatic speech recognition task has been presented. Used toolkits, libraries and prepared speech corpus have been described. The obtained results suggest, that using different acoustic models for normal speech and whispered speech can reduce word error rate. The optimal training steps has been also selected. Thanks to the additional simulations it has been found that used corpus (over 9 hours of normal speech and the same of the whispery speech) is definitely too small and must be enlarged in the future.*

***Streszczenie.** W artykule przedstawiono automatyczne rozpoznawanie mowy. Wykorzystane narzędzia, biblioteki i korpus opisano w artykule. Uzyskane wyniki wskazują, że wykorzystując różne modele akustyczne dla mowy zwykłej i szeptanej uzyskuje się polepszenie skuteczności rozpoznawania mowy. W wyniku wykonanych badań wskazano również optymalną kolejność kroków treningu. Dzięki dodatkowym obliczeniom stwierdzono, że użyty korpus (ponad 9 godzin zwykłej mowy i drugie tyle szeptu) jest zdecydowanie za mały do dobrego wytrenowania systemu rozpoznawania mowy i w przyszłości musi zostać powiększony. (**Narzędzia Kaldi w rozpoznawaniu polskiej mowy szeptanej**).*

Keywords: speech recognition, ASR, whispered speech, database.
Słowa kluczowe: rozpoznawanie mowy, ASR, mowa szeptana, baza danych.

Introduction

The automatic speech recognition (ASR) systems become more widely used and are available in most of modern phones as well as in many websites. Those systems are, usually, an alternative to the manual text input, e.g. SMS messages. ASR can be also used for more sophisticated tasks, as support of a speech transcription (e.g. in a parliament or in an army).

Most of research in the literature is focused on a normal speech, while the whispery speech is rarely explored (but the largest electronics companies are interested in this topic [1,2]). Apart from the military and reconnaissance usage, automatic whispery speech recognition (AWSR) systems can be useful for people who are unable to speak normally, e.g. after laryngectomy [3].

An appropriate speech corpus is necessary for the research in AWSR. One of such databases is NAM TIMIT Plus, in which there are about one thousand whispered utterances. A little bit more one can find in CHAINS database [4] – about 1200 sentences (corpus is larger, but the whispery speech is only a part of this corpus). It is much less in comparison to normal speech corpuses (e.g. the AGH corpus contains 25 hours of recordings [5]). Hence, a new corpus has been prepared for the research.

In the first section the architecture of the ASR system used in the reported experiments has been described. Section 2 contains a description of the tools which have been used together with Kaldi toolkit. A new speech corpus, prepared by the authors has been presented in Section 3. Next, the results of the performed experiments have been shown in the following section. The conclusions one can find in Section 5.

Speech re

The automatic speech recognition system performs conversion from an audio signal to a text. Hidden Markov models (HMMs) are usually used for this task [6]. HMMs can model stochastic processes, in which observations are generated by hidden states. There are transition probabilities between these states [7].

Before HMM the feature vector must be received. In each time frame (very short, 10-30ms) the speech signal is converted into a feature vector. The features extraction steps are [8]:

- preemphasis (initial filtration of the input signal),
- division into frames (number of samples in a frame should be a power of 2),
- windowing (Kaldi allows easy usage of 4 windowing types, including one proposed by toolkit authors),
- fast Fourier transform (FFT),
- transition into the Mel scale,
- decorrelation of the elements of the feature vector by means of discrete cosine transform (DCT)

For each frame, these steps result in twelve values, which are supplemented by signal energy. For these 13 feature values the first and second derivatives are calculated (Δ+ΔΔ). Finally, the obtained vector, composed of 39 values, describes spectral envelope and dynamics of its change.

A creation and appropriate training of the model (composed of, e.g., Gaussian mixture models – GMM) also must be done. Such model allows conversion from feature vectors into specific phones. Speech models (acoustic – HMM; lexicon – dependences between phones and words; language) allow to obtain the most probable word, or words sequence, based on the speech signal. It is done in the decoding step, in which the most feasible sentence $\hat{w}$ must be found, using the observation (vector of features) sequence $O=\{o_1,o_2,...,o_M\}$, language model $p(w)$ and acoustic model $p(O|w)$. It can be written as

$$\hat{w} = \arg\max_{w}\left(p(w \mid O)\right) = \arg\max_{w}\left(\frac{p(w)p(O \mid w)}{p(O)}\right) = \\ = \arg\max_{w}\left(p(w)p(O \mid w)\right). \tag{1}$$

It should be noted that GMMs and HMMs are the most commonly used in ASR task, but the approaches with neural networks become increasingly prevalent [9,10].

Used toolkits and libraries

The Kaldi is the most important of used tools. It has been written in C++ and is licensed under the Apache v2.0 [11]. The training of acoustic model (AM) in Kaldi is composed of few steps. The first step (mono) uses monophones – this step usually is used only as the initialization of the recognition model. In the second step (tri1) triphones are used, i.e. the three subsequent phones are taken into account. In the next step (tri2a) AM model is ready, based on the all 39 features and using triphones. Another approach is the tri2b step, in which the Linear Discriminant Analysis (LDA) and Maximum Likelihood

Linear Transform (MLLT) have been used to transform Mel Frequency Cepstral Coefficients (MFCC) features [12]. At the last step (sgmm2) the subspace Gaussian mixture models (SGMMs) are used in AM model.

Two external libraries have been used in Kaldi, i.e. BLAS/LAPACK (available on the website www.netlib.org) to perform linear algebra computations, and OpenFST [13], which allows an efficient application of the finite state transducers (FST). All models (acoustic, lexicon and language) can be represent by weighted FST (WFST).

Additionally the SRILM package has been used [14], which is used to a preparation of a language model based on the available data (e.g. using transcriptions). Such model contains the information about possible connections between words in a sentence (e.g. the words sequence “train flying under water” is rather not very likely; hence, even if such connections will occur, it will occur with very small probability).

The last program, which has been used is Sequitur [15], which allows conversion graphemes→phonemes (G2P), i.e. replaces words to phonemes sequence. This tool operates independently of the language; however, it requires the creation of the G2P model based on some prior knowledge.

Created speech corpus is in Polish, so the G2P model for Polish has been trained. The Wiktionary contains a pronunciation of presented words and this information has been used as the prior knowledge – all entries from Polish Wiktionary have been downloaded and the pronunciation in IPA (International Phonetic Alphabet) system has been used. However, one should keep in mind that IPA system contains allophones, and not phonemes (one phoneme can has few different phones representation). Hence 87 allophones have been modeled (number of Polish allophones), and not 39 phonemes (typically for Polish system SAMPA) [16], what can has impact on the obtained results.

Speech corpus

The created database is composed of recordings carried out by 33 persons, aged 20 to 25 years (students). The speech corpus contains both normal and whispery speech. In the Table 1 the corpus properties have been shown in numbers.

Table 1. Properties of the used speech corpus

Speech corpus property	Normal speech	Whispery speech
Number of sentences	5935	5411
Number of words	61964	53335
Number of different words	3556	3427
Total recordings length [min]	547.8	548.5

All recordings have been done by speakers on their own hardware, therefore obtained audio signals differ in both voice quality and noise level. All recordings have been saved in 16-bit and 48 kHz.

Contents of utterances have been taken from the website wolnelektury.pl, where literary works are available in the public domain (used works include “Seasonal Love”, “Snow-White and Rose-Red”, “The Ugly Duckling”, “The Fir Tree”, “The Toad”, “The Nightingale”, and others).

Obtained results

Due to a small amount of data, the training utterances for each speaker were different, in such a way that training were performed based on all other speakers. Simulations have been done for different types of training and test speech (normal, whispery and both). In calculations two method paths have been used: mono → tri1 → tri2a and mono → tri1 → tri2b → ubm → sgmm2 (designations from Kaldi toolkit). Obtained results have been presented in Table 2 (all results are for the 2-gram language model).

The time needed for training (without decoding) was appropriately 7 min (data preparation), 5 min (mono), 10 min (tri1), 10 min (tri2a), 11 min (tri2b), 170 min (sgmm2). Decoding in most of cases takes 10 min, and the exceptions are sgmm2 (18 min) and sgmm2 with fmllr option (27 min).

The additional simulations have been performed using 1-gram (mono-gram) language model (LM), and also simulations in which the whole available data have been used during training (including recordings, which were used for testing). The results have been presented in the figures 1-2 (only Word Error Rate %WER after tri2a and sgmm2+fmllr steps). Each presented value is the mean of three speakers (no. 8, 21 and 29) results. The higher value in each bar is associated with the training without the additional data, and the lower value (in brackets) is for training, in which also the testing data has been used.

Table 2. Results obtained for second-order language model (bi-grams have been used); presented values means %WER, i.e. the percent of mistakenly recognized words

	test speech	normal	normal	normal	whispery	whispery	whispery
Speaker	training speech	normal	whispery	both	normal	whispery	both
8	mono	33.08	58.88	34.85	51.95	50.23	45.65
	tri1	26.94	44.79	25.70	44.19	42.29	35.57
	tri2a	26.78	43.60	25.75	43.90	41.48	36.16
	tri2b	26.94	63.46	25.66	48.33	42.76	35.62
	sgmm2	24.97	54.18	24.33	44.76	40.96	33.59
	sgmm2+fmllr	24.60	51.59	23.55	43.20	40.18	32.91
29	mono	12.27	37.57	18.13	48.32	31.47	40.51
	tri1	5.15	25.18	7.06	37.09	20.31	22.65
	tri2a	5.01	24.71	7.41	35.86	18.35	21.74
	tri2b	5.89	27.44	6.24	40.48	20.98	17.69
	sgmm2	3.63	20.97	5.89	36.55	14.54	18.08
	sgmm2+fmllr	2.96	18.51	5.12	33.68	13.83	16.41
21	mono	40.17	70.50	46.83	89.11	67.48	76.43
	tri1	26.01	53.27	28.55	85.58	54.57	56.09
	tri2a	28.68	51.02	25.21	85.41	52.97	55.25
	tri2b	18.91	65.45	20.40	91.78	75.22	53.62
	sgmm2	18.75	65.15	24.69	93.21	69.89	57.77
	sgmm2+fmllr	13.93	59.77	19.27	92.70	67.03	53.31

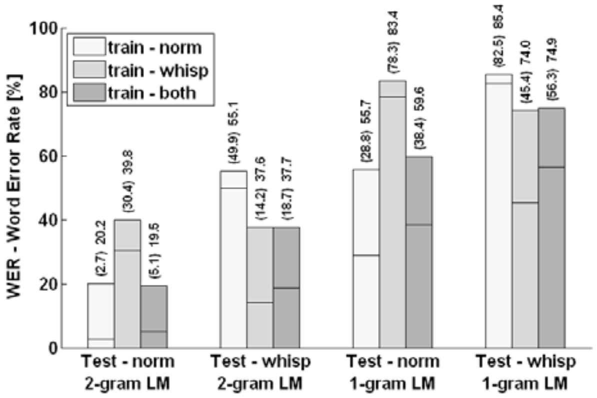

Fig.1. Results obtained after tri2a step in Kaldi; presented values are the means of three speakers results; the lower value (in brackets) is associated with training, in which the testing data has also been used, and the higher value is for training without testing data.

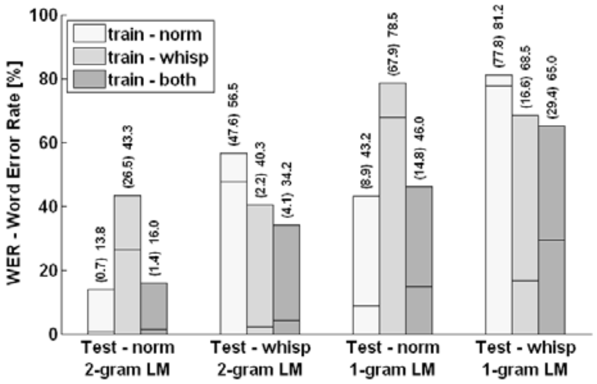

Fig.2. Results obtained after sgmm2 step (with fmllr option) in Kaldi; presented values are the means of three speakers results; the lower value (in brackets) is associated with training, in which the testing data has also been used, and the higher value is for training without testing data.

Conclusions

The most important conclusion, which comes from obtained results is that creation of separate ASR models for normal and whispery speech makes sense (comparing the results, where test and training were for the same speech type, with the results, where training was performed for both speech types) – for whispery speech the results are approximately the same, but with two times larger training set the recognition rate should be noticeably better.

Additionally, based on the obtained results and calculations time it has been found that for the future research the best training path is mono → tri1 → tri2a, which allows to obtain satisfactory results in a relatively short time period (6 times faster than for training to the sgmm2).

It is worth paying attention to results of speaker no. 29 for normal speech during training and test – the recognition rate was over 97% (only 3% Word Error Rate), what shows that the usage of all allophones is not a fundamental mistake. In the article, which has been accepted on the MMAR 2016 conference, it has been presented that changes of few allophones may have positive influence on the word recognition rate.

On the other hand, looking on the same type of training and testing speech (normal/normal) for the speaker No. 8, %WER was over eight times higher (24,6%). Generally, one can see high differences in all results for different speakers. It is caused by the fact that each speaker had used his own device for recordings, and differences in the noise level (see Fig. 3) had a significant impact on the obtained results.

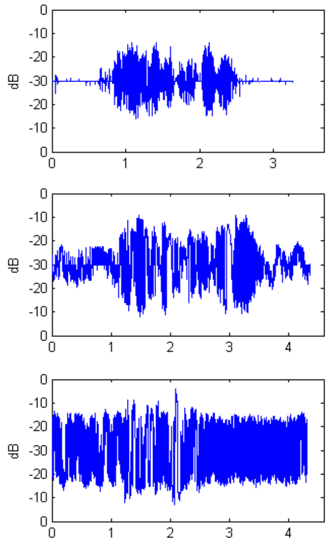

Fig.3. The comparison of waveforms for sentence “Prześlicznie było na wsi.”, which were obtained from 3 different speakers; waveforms are presented in specific logarithmic scale, where upper -10dB means signal value +0.1, and lower -10dB means signal value -0.1; all values in range from -0.001 to 0.001 are presented at the same level: -30dB

Based on the results in figures 1-2 one can say that the used speech corpus is too small, because the difference between training with and without testing data should not be so large. The speech corpus should be expanded as long as the difference will be noticeable.

One can also say that the normal and whispery speech have some joint features, because without that there should not be the improvement after adding the testing data (cases with different speech type in training end testing). When such feature coefficients will be found, the one model should be sufficient for both normal and whispery speech.

The authors plan to increase database in the future. Except that the research on the whispery speech will be deepened, including finding the best coefficients set, which will provide as good recognition quality as in the normal speech.

Artykul zostal wygloszony w ramach Lubelskich Dni Nauki i Biznesu w Lublinie na konferencji naukowej 11-13 Czerwca WD 2016 w ramach warsztatów doktoranckich organizowanych przez Centrum Badawczo Rozwojowe NETRIX SA pod patronatem honorowym Polskiego Stowarzyszenia Tomografii Procesowej i Komitetu Elektrotechniki PAN przy współpracy między innymi z Instytutem Elektroniki i Technik Informacyjnych Politechniki Lubelskiej i Instytutu Elektrotechniki w Warszawie.

***Authors**: mgr inż. Piotr Kozierski, Poznan University of Technology, Faculty of Computing, Chair of Control and Systems Engineering, Division of Signal Processing and Electronic Systems and Faculty of Electrical Engineering, Institute of Control and Information Engineering, Division of Control and Robotics, Piotrowo 3a street, 60-965 Poznań, E-mail: piotr.kozierski@gmail.com; mgr inż. Talar Sadalla, Poznan University of Technology, Faculty of Electrical Engineering, Institute of Control and Information Engineering, Division of Control and Robotics, E-mail: talar.h.sadalla@doctorate.put.poznan.pl; dr Szymon Drgas, Poznan University of Technology, Faculty of Computing, Chair of Control and Systems Engineering, Division of Signal Processing and Electronic Systems, Email: szymon.drgas@put.poznan.pl; prof. dr hab. inż. Adam Dąbrowski, Poznan University of Technology,* *Faculty of Computing, Chair of Control and Systems Engineering, Division of Signal Processing and Electronic Systems, Email: adam.dabrowski@put.poznan.pl; dr hab. inż. Dariusz Horla, Poznan University of Technology, Faculty of Electrical Engineering, Institute of Control and Information Engineering, Division of Control and Robotics, Email: dariusz.horla@put.poznan.pl.*

REFERENCES

[1] Hong S.J., Method and Apparatus for Recognizing Whisper, *U.S. Patent Application*, No. US14579134 (filed December 22, 2014)

[2] Huang X., Acero A., Alleva F., Hwang M.Y., Jiang L., Mahajan M., Microsoft Windows Highly Intelligent Speech Recognizer: Whisper, *In Acoustics, Speech, and Signal Processing, 1995 International Conference on* (ICASSP-95), 1 (May 1995), 93-96

[3] Sharifzadeh H.R., McLoughlin I.V., Ahmadi F., Reconstruction of Normal Sounding Speech for Laryngectomy Patients through a Modified CELP Codec, *Biomedical Engineering, IEEE Transactions on*, 57 (2010), No. 10, 2448-2458

[4] Cummins F., Grimaldi M., Leonard T., Simko J., The Chains Corpus: Characterizing Individual Speakers, *In Proc. of SPECOM*, 6 (2006), 431-435

[5] Żelasko P., Ziółko B., Jadczyk T., Skurzok D., AGH Corpus of Polish Speech, *Language Resources and Evaluation*, (2015), 1-17, DOI: 10.1007/s10579-015-9302-y

[6] Szostek K., Optimization of HMM models and their application in speech recognition (in Polish), *Elektrotechnika i Elektronika*, 24 (2005), No. 2, 172-182

[7] Plannerer B., An Introduction to Speech Recognition, Munich, Germany (2005)

[8] Wanat I., Iwaniec M., Creation of the acoustic model for speaker recognition using hidden Markov models (in Polish), *Modelowanie Inżynierskie*, 9 (2010), No. 40, 249-256

[9] Miao Y., Kaldi+PDNN: Building DNN-based ASR Systems with Kaldi and PDNN, *arXiv preprint* arXiv:1401.6984 (2014)

[10] Mohanty R., Mohanty P., A Review: Neural Networks used for Speech Recognition, *IJRECE*, 4 (2016), No. 1, 01-05

[11] Povey D., Ghoshal A., Boulianne G., Burget L., et al., The Kaldi Speech Recognition Toolkit, *In IEEE 2011 workshop on automatic speech recognition and understanding*, (2011), No. EPFL-CONF-192584

[12] Platek O., Speech Recognition using KALDI, *Master thesis*, Charles University in Prague, Faculty of Mathematics and Physics (2014)

[13] Allauzen C., Riley M., Schalkwyk J., Skut W., Mohri M., OpenFst: A General and Efficient Eeighted Finite-State Transducer Library, *In Implementation and Application of Automata*, Springer Berlin Heidelberg (2007), 11-23

[14] Stolcke A., SRILM-an Extensible Language Modeling Toolkit, *In Proc. Intl. Conf. Spoken Language Processing* (INTERSPEECH), Denver, Colorado (September 2002)

[15] Bisani M., Ney H., Joint-Sequence Models for Grapheme-to-Phoneme Conversion, *Speech Communication*, 50 (2008), No. 5, 434-451

[16] Wypych M., Baranowska E., Demenko G., A Grapheme-to-Phoneme Transcription Algorithm Based on the SAMPA Alphabet Extension for the Polish Language, *Phonetic Sciences, 15th International Congress of* (ICPhS), Barcelona (August 2003), 2601-2604