Position Papers of the Federated Conference on Computer Science and Information Systems pp. 109–114 DOI: 10.15439/2018F255 ISSN 2300-5963 ACSIS, Vol. 16

Acoustic Model Training, using Kaldi, for Automatic Whispery Speech Recognition

Piotr Kozierski[1,2)], Talar Sadalla[2)], Szymon Drgas[1)], Adam Dąbrowski[1)], Joanna Ziętkiewicz[2)], Wojciech Giernacki[2)]
Poznan University of Technology, Piotrowo street 3a, 60-965 Poznan, Poland[1,2)]
Faculty of Computing, Institute of Automation and Robotics, Division of Signal Processing and Electronic Systems[1)]
Faculty of Electrical Engineering, Institute of Control, Robotics and Information Engineering, Division of Control and Robotics[2)]
Email: piotr.kozierski@gmail.com, szymon.drgas@put.poznan.pl

***Abstract*—The article presents research on the automatic whispery speech recognition. The main task was to find dependences between a number of triphone classes (number of leaves in decision tree) and the total number of Gaussian distributions and therefore, to determine optimal values, for which the quality of speech recognition is best. Moreover, it was found, how these dependences differ between normal and whispery speech, what was not done earlier, and this is the innovative part of this work. Based on the performed experiments and obtained results one can say that the number of triphone classes (number of leaves) for whispered speech should be significantly lower than for normal speech.**

I. Introduction

WHISPERS are relatively rarely used in comparison to normal speech. Usually people whisper in specific environment or during private communication [1]. However, for persons after laryngectomy operation, the whispered speech is the only way to communicate with others without special prosthesis [2].

The largest companies (such as Microsoft or Apple) are interested in whispered speech recognition [3]-[4], and also in military domain one of research directions is focused on Automatic Speech Recognition (ASR) systems [5]-[6].

One can find studies on ASR systems for whispered speech [7], even a whispering speaker identification [8]; however, there is still very little research in this area. And even if some studies are provided about whispered speech, very small corpora are used (in latter two references corpora contain less than 500 sentences in sum).

In this paper the authors were focused on the acoustic model training. The most common approach, in which Gaussian Mixture Models (GMMs) are used for Probability Density Functions (PDFs) of features vector values modeling, was taken into account; however, one can find also other approaches, such as Decision Tree-based Acoustic Models (DTAM), in which decision trees are used instead of of GMMs [9].

The research task of this paper was to find the optimal value of Gaussian distributions for the given number of leaves, and the optimal number of triphone classes for given number of Gaussian distributions simultaneously. Moreover, the differences between normal and whispery speech were investigated, because such research has not been performed before.

In the second section, one can find a description of software which was used during studies. Information about operation principle of automatic speech recognition are given in Section III. Section IV contains details about speech corpus, which was used in research. In fifth section, one can find description of quality index and obtained results. In the last section, drawn conclusions are presented.

II. Used Software – Kaldi

During studies the Kaldi toolkit [10] was used, which contains scripts and programs for speech recognition task. This software is available under Apache v2.0 license, is easy to change and is still being developed. In Kaldi, two external libraries are used, i.e. BLAS/LAPACK for linear algebra calculations (library is available on the website www.netlib.org) and OpenFST [11]. The latter is used due to the fact that in Kaldi Finite State Transducers (FSTs) are used as representation of most of data [12].

Moreover, SRILM package [13] and Sequitur [14] programs were used. The first one was used for Language Model (LM) preparation (including Witten-Bell smoothing [15]) – it contains information, how .

The second one was used for graphemes-to-phonemes (G2P) conversion – the tool is language independent. The studies were performed for Polish, hence the G2P model was trained (based on the International Phonetic Alphabet – IPA – pronunciation available in Wiktionary). Extended SAMPA notation [16]-[17] with 39 phonemes was used instead of SAMPA notation with 37 phonemes [18]. For comparison, there are 55 phonemes in Russian and 49 phonemes in American English [19].

SAMPA (Speech Assessment Methods Phonetic Alphabet) is a standard for describing specific phonemes

This work was not supported by any organization

(and allophones) by signs from 7-bit ASCII. In this approach it is assumed that there are 37 phonemes (sounds) in Polish.

Extended SAMPA notation was proposed by authors to improve speech recognition of Polish, hence that standard was used in performed research. Differences between SAMPA and Extended SAMPA notations are presented in Table I (for whole list of Polish phonemes see [16] or [20]).

III. Automatic Speech Recognition

The main task of ASR systems is to decode the most probable word sequence (or only word, if isolated words are recognized instead of continuous speech), based on the audio signal with speech. The audio signal is divided into very short (usually 16-25 ms [21], but 25 ms default value was used) overlapping parts (the frame shift default value was equal to 10 ms), so called frames. From each frame, a feature vector is obtained – for Mel Frequency Cepstral Coefficients (MFCC) it is usually 13 values (twelve based on windowing, FFT, Mel scaling and DCT transformations and 13-th – signal energy). Next, first and second derivatives are calculated (Δ+ΔΔ) obtaining 39 MFCC features.

Each feature vector is associated with i-th signal frame and is treated as observation o_i. The whole recording is a sequence of such observations $O=\{o_1, o_2, ..., o_M\}$. Among all possible utterances (word sequences) w one must find the most likely sentence, and it can be written as

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

where $p(w|O)$ is a posterior PDF, which must be maximized, $p(w)$ is the prior PDF, which informs, what is the probability that sentence w occurs, $p(O|w)$ is the conditional probability that word sequence w occurs for observations O, and $\hat{w}$ is the most probably words sequence.

Table I.
Differences between SAMPA and Extended SAMPA notations [16]

Word in Polish	SAMPA		Ext. SAMPA	
	notation	**transcription**	**notation**	**transcription**
ty̲p	I	t **I** p	y	t **y** p
gęś	e~	g **e~** s’	–	–
wąs	o~	v **o~** s	–	–
kat	k	**k** a t	k	**k** a t
gen	g	**g** e n	g	**g** e n
kiedy	–	–	c	**c** j e d y
giełda	–	–	J	**J** j e w d a
cyk	ts	**ts** I k	t^s	**t^s** y k
dzwon	dz	**dz** v o n	d^z	**d^z** v o n
czyn	tS	**tS** I n	t^S	**t^S** y n
dżem	dZ	**dZ** e m	d^Z	**d^Z** e m
ćma	ts’	**ts’** m a	t^s’	**t^s’** m a
dźwig	dz’	**dz’** v i k	d^z’	**d^z’** v i k
ciąża	–	–	w~	t^s’ o **w~** Z a
więź	–	–	j~	v j e **j~** s’

The whole ASR model is constructed as the Hidden Markov Model (HMM), in which each state is associated with some phoneme (or triphone). Based on the subsequent observations, ASR system should estimate sequence of these states, and this can be solved using the Viterbi algorithm [22].

The prior probability $p(w)$ is represented by the LM in ASR model. Similarly, Acoustic Model (AM) represents probability $p(O|w)$. Therefore, one can see that appropriate LM and AM are fundamental for good ASR working. Language model contains information about “word connections”, one can say “grammar” of specific language. Very helpful may be corpus which contains a huge number of works (articles, novels, poems, blog entries, etc.), e.g. [23]. However, in practice, only utterances available in speech corpus are taken into account.

Preparation of LM is relatively fast. Much more difficult is AM training. Acoustic model is also constructed in HMM form; hence, transition probabilities are calculated during training, but also probabilistic distributions associated with specific states (triphones) must be estimated. These distributions are usually represented by the Gaussian mixture models; however, one can find also other approaches [9]. GMM is a distribution which is created from the combination (addition) of two or more Gaussian distributions.

The all states (associated with triphones – three consecutive phonemes) are clustered and not all possible triphones are modeled. It does not mean that different states are treated as the same, but at this processing stage few (or dozen) triphones from the same class are unrecognizable. Based on the lexicon or LM a specific triphone will be recognized later. In Kaldi toolkit this classification is done by the Decision Tree (DT), i.e., one specific class is chosen based on the series of comparisons (which parameters are compared and boundary values are chosen during AM training).

The size of DT (number of classes) and the whole number of Gaussian distributions (each GMM, which described one class, is composed of few or dozen ones) are the main two parameters in AM training in Kaldi.

IV. Speech Corpus

The authors were focused on the Automatic Whispery Speech Recognition (AWSR), and hence a specific speech corpus was needed. However there are very few of databases with whispery speech. One of such is CHAINS corpus [24] which contains about 1200 sentences in whisper. The second available corpus – Audiovisual (AVW) [25] contains over 1300 whispered sentences. Unfortunately, both are too small for AWSR research.

The largest corpus with whispery speech, which was used in [26], contains about 14,000 whispered sentences and theoretically could be used in AWSR task; however, this is Japanese corpus, and in such languages like Mandarin or Japanese it is important to model the accent, which can change the word meaning [27].

Due to the lack of required database, the authors decided to prepare corpus with Polish (normal and whispered) speech (see Table II, and for comparison also older version of corpus is described in Table III). The sentences come mainly from Andersen's fairy tales (like "The Toad", "The Nightingale", "The Ugly Duckling"); however, one can find also fragments from Grimm brothers' fairy tales.

All utterances were recorded in 48 kSps sampling rate and 16-bit quantization depth. Every speaker (there are over 50 different speakers) recorded sentences on his/her own device, so the recordings quality widely vary between speakers.

V. EXPERIMENTS AND RESULTS

All experiments were repeated two times for different test speakers group (the choice of testing speakers was widely described in [28]) and presented results are mean values. Each time ASR system was trained on recordings from all speakers (except the testing ones).

The quality of speech recognition is described by the Word Error Rate (WER) index

$$\mathrm{WER} = \frac{\mathrm{Del} + \mathrm{Ins} + \mathrm{Subs}}{N_{utt}} \cdot 100\% \,, \qquad (2)$$

where Del is the number of deletions (cases where word from reference is not present in output sentence), Ins is the number of insertions (cases where word in recognized sentence does not occur in reference), Subs is the number of substitutions (cases where one word from reference is confused with another one from output sentence), and N_{utt} is the number of words in a reference sentence. The sum in numerator is a minimum edit distance on words between obtained output from ASR system and the reference utterance [12].

TABLE III.
PROPERTIES OF THE USED SPEECH CORPUS (2^{ND} VERSION)

Property	Normal speech	Whispered speech
Number of sentences	9,522	8,753
Number of words	108,038	95,305
Number of different words	5,094	4,763
Total recordings length	988.5 min (16.5 h)	942.9 min (15.7 h)
Number of speakers	56	

TABLE III.
PROPERTIES OF THE OLD SPEECH CORPUS (1^{ST} VERSION) – ALL THESE RECORDING ARE ALSO CONTAINED IN 2^{ND} VERSION

Property	Normal speech	Whispered speech
Number of sentences	5,935	5,411
Number of words	61,964	53,335
Number of different words	3,556	3,427
Total recordings length	547.5 min (9.1 h)	548.5 min (9.1 h)
Number of speakers	33	

The lexicon was extended from 5,000 words, which occur in speech corpus, to 50,000 words to obtain large vocabulary ASR system (based on the classification in [29]). It was done by adding new sentences during LM creation. Moreover, the differences in results and directions of changes are better visible for higher WER level.

During research the AM training path mono → tri1 → tri2a (designations from Kaldi) was used. The choice was dictated by the previous studies [30], where it was concluded that this training path provides satisfying speech recognition quality and quite short training computation time simultaneously.

The experiments were performed for different number of leaves and number of Gaussian distributions in AM training. In each case values were the same for tri1 and tri2a steps – it was caused by the preliminary research. In both tri1 and tri2a steps the same scripts are used, and the only difference between tri1 and tri2a is their order. The authors tried to add third step tri3a (again – the same script run third time) and the impact of leaves number and Gauss number in first two runs on speech recognition quality was studied (in all cases numbers of leaves and Gauss were the same for third – tri3a – step).

The results of preliminary research showed that only parameters in the last step have significant influence on the speech recognition quality. Based on this, it was decided to set the same parameter values for both tri1 and tri2a steps. Obtained results are presented in Fig. 1-4.

VI. CONCLUSIONS

Based on the obtained results presented in Fig. 1-2 one can see that with increasing the total number of Gaussian distributions in AM model the speech recognition quality improves. This is caused by better modeling of the probability distribution of phones' features vectors (more sum components in GMMs).

However, for the number of leaves (number of different triphones classes – see Fig. 3-4) one can see that there is an optimal value, for which WER index is the lowest. For normal speech, number of Gaussian distributions should be increased together with the number of classes. And the default values (2,000 leaves and 11,000 Gaussian distributions) are quite good for normal speech (if one wants to improve acoustic model, and quality of ASR system at the same time, both values should be increased).

On the contrary, in whispery speech one can see the difference – the optimal value of decision tree leaves is lower in all cases. The default value (2000) should be decreased, or the number of Gaussian distributions should be greatly increased; however, in all cases, further increase of class number has a negative impact on speech recognition quality. This is probably caused by the fact that in whispered speech there is no more than 2,000 different triphones (theoretically it could be $39^3 \approx 60,000$; however, in whispers many phonemes sounds the same).

In the future research the authors plan to use neural networks for whispery speech recognition, and also sharing of prepared speech corpus is planned.

References

[1] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, “Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec,” *Biomedical Engineering, IEEE Transactions on*, vol. 57, no. 10, pp. 2448–2458, 2010.

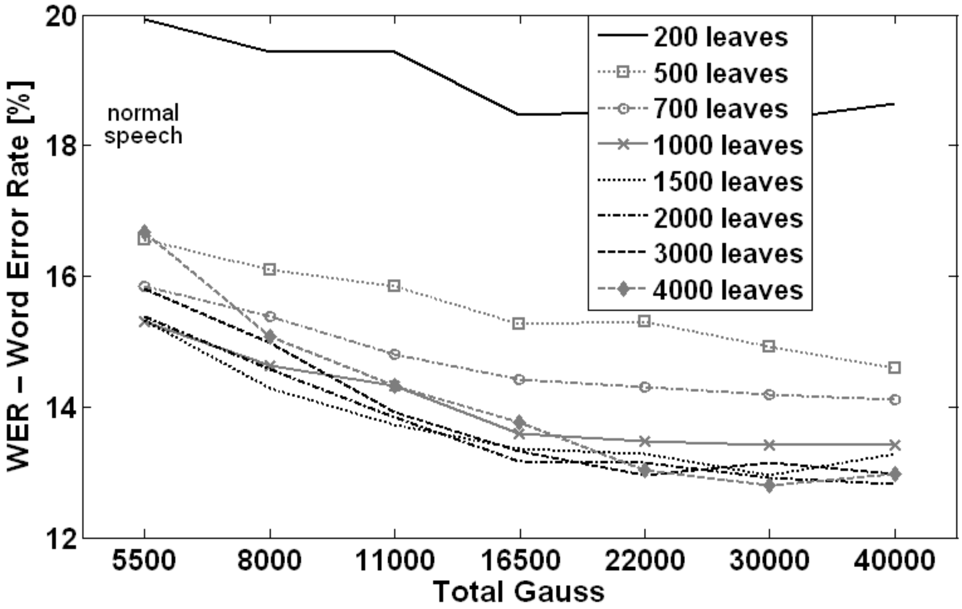

Fig. 1 Quality of speech recognition in function of total Gaussian distributions number for normal speech

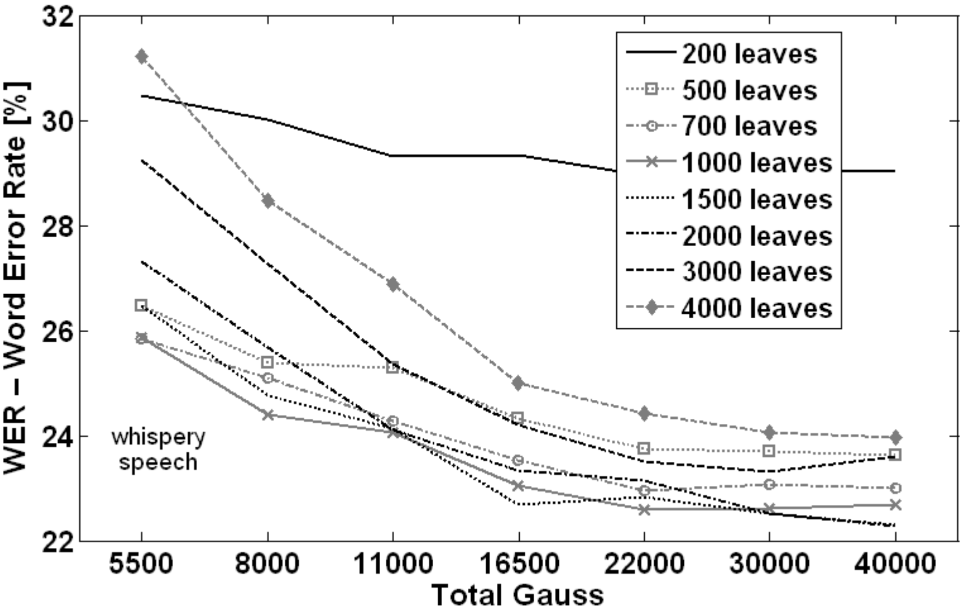

Fig. 2 Quality of speech recognition in function of total Gaussian distributions number for whispery speech

[2] H. F. Nijdam, A. A. Annyas, H. K. Schutte, and H. Leever, "A new prosthesis for voice rehabilitation after laryngectomy," *Archives of Otorhinolaryngology*, vol. 237, no. 1, pp. 27–33, 1982.

[3] X. Huang, A. Acero, F. Alleva, M. Y. Hwang, L. Jiang, and M. Mahajan, "Microsoft Windows highly intelligent speech recognizer: Whisper," in *Acoustics, Speech, and Signal Processing, 1995 International Conference on (ICASSP-95)*, vol. 1, pp. 93–96.

[4] T. J. Raitio, M. J. Hunt, H. B. Richards, and M. Chinthakunta, "Digital assistant providing whispered speech," U.S. Patent 15/266,932, December 14, 2017.

[5] D. T. Williamson, M. H. Draper, G. L. Calhoun, and T. P. Barry, "Commercial speech recognition technology in the military domain: Results of two recent research efforts," *International Journal of Speech Technology*, vol. 8, no. 1, pp. 9–16, 2005.

[6] S. Pigeon, C. Swail, E. Geoffrois, G. Bruckner, D. Van Leeuwen, C. Teixeira, et al., *Use of speech and language technology in military environments*, Montreal, Canada, North Atlantic Treaty Organization, 2005.

[7] S. C. S. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *ICASSP*, March 2005, pp. 1009–1012.

[8] Q. Jin, S. C. S. Jou, and T. Schultz, "Whispering speaker identification," in *Multimedia and Expo, 2007 IEEE International Conference on*, pp. 1027–1030.

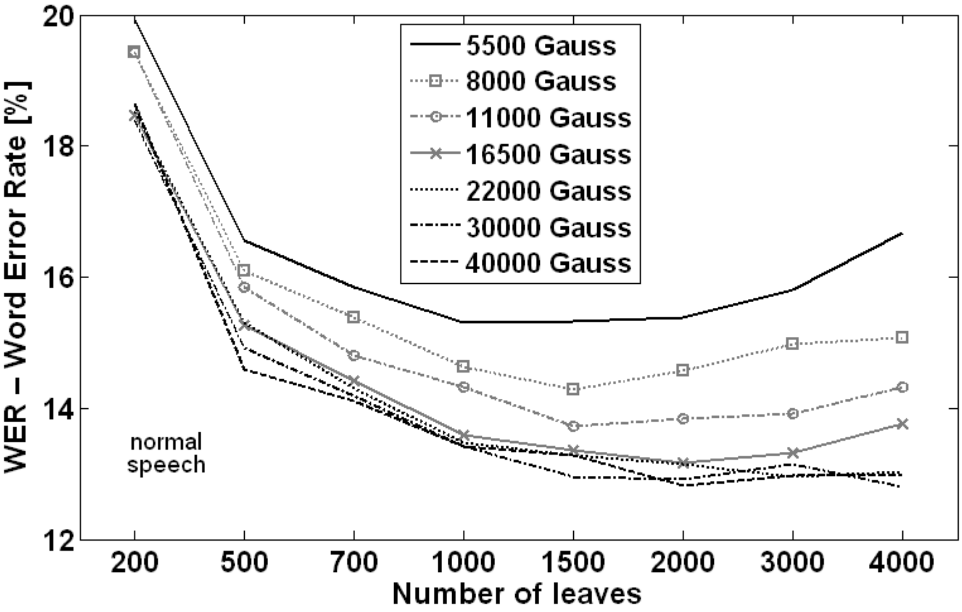

Fig. 3 Quality of speech recognition in function of leaves number for normal speech

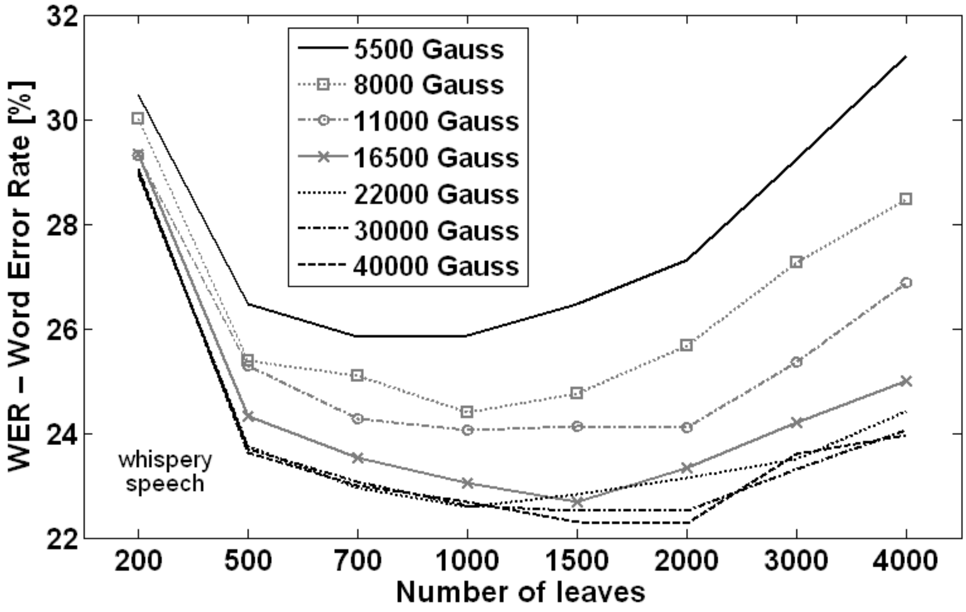

Fig. 4 Quality of speech recognition in function of leaves number for whispery speech

[9] M. Akamine, and J. Ajmera, "Decision tree-based acoustic models for speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2012, art. no. 10, p. 8, 2012.

[10] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, No. EPFL-CONF-192584, 2011.

[11] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *Implementation and Application of Automata*, J. Holub and J. Žďárek, Ed. Berlin: Springer Heidelberg, 2007, pp. 11–23.

[12] O. Platek, "Speech recognition using KALDI," M.S. thesis, Inst. Form. Appl. Ling., Charles Univ., Prague, Czech Republic, 2014.

[13] A. Stolcke, "SRILM-an extensible language modeling toolkit," in *Proc. Intl. Conf. Spoken Language Processing (INTERSPEECH)*, Denver, Colorado, September 2002, pp. 901–904.

[14] M. Bisani, and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech Communication*, vol. 50, no. 5, pp. 434–451, 2008.

[15] I. H. Witten, and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," *IEEE Transactions on Information Theory*, vol. 37, no. 4, pp. 1085–1094, 1991.

[16] G. Demenko, M. Wypych, and E. Baranowska, "Implementation of grapheme-to-phoneme rules and extended SAMPA alphabet in Polish text-to-speech synthesis," *Speech and Language Technology*, vol. 7, pp. 79–97, 2003.

[17] M. Wypych, E. Baranowska, and G. Demenko, "A grapheme-to-phoneme transcription algorithm based on the SAMPA alphabet extension for the Polish language," in *Phonetic Sciences, 15th International Congress of (ICPhS)*, Barcelona, August 2003, pp. 2601–2604.

[18] P. Kłosowski, "Improving speech processing based on phonetics and phonology of Polish language," *Przeglad Elektrotechniczny*, vol. 89, no. 8, pp. 303–307, 2013.

[19] A. Karpov, K. Markov, I. Kipyatkova, D. Vazhenina, and A. Ronzhin, "Large vocabulary Russian speech recognition using syntactico-statistical language modeling," *Speech Communication*, vol. 56, pp. 213–228, 2014.

[20] P. Kozierski, T. Sadalla, S. Drgas, A. Dąbrowski, "Allophones in automatic whispery speech recognition," *in Methods and Models in Automation and Robotics (MMAR), 21st International Conference on*, 2016, pp. 811-815. DOI: 10.1109/MMAR.2016.7575241

[21] F. Portet, M. Vacher, C. Golanski, C. Roux, and B. Meillon, "Design and evaluation of a smart home voice interface for the elderly: Acceptability and objection aspects," *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.

[22] K. Szostek, "Optimization of HMM models and their usage in speech recognition (in Polish)," *Elektrotechnika i Elektronika*, vol. 24, no. 2, pp. 172–182, 2005.

[23] B. Lewandowska-Tomaszczyk, M. Bańko, R. L. Górski, P. Pęzik, and A. Przepiórkowski, *National corpus of Polish language (in Polish)*, Warszawa: Wydawnictwo Naukowe PWN, 2012.

[24] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The Chains corpus: Characterizing individual speakers," *in Proc. of SPECOM*, vol. 6, 2006, pp. 431–435.

[25] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 8101–8105.

[26] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[27] C. Huang, E. Chang, J. Zhou, K. and F. Lee, "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition," in *INTERSPEECH*, October 2000, pp. 818–821.

[28] P. Kozierski, T. Sadalla, S. Drgas, A. Dąbrowski, and J. Zietkiewicz, "The impact of vocabulary size and language model order on the Polish whispery speech recognition," in *Methods and Models in Automation and Robotics (MMAR), 22nd International Conference on*, 2017, pp. 616–621. DOI: 10.1109/MMAR.2017.8046899

[29] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Communication*, vol 56, pp. 85–100, 2014.

[30] P. Kozierski, T. Sadalla, S. Drgas, A. Dąbrowski, and D. Horla, "Kaldi toolkit in Polish whispery speech recognition," *Przeglad Elektrotechniczny*, vol. 92, no. 11, pp. 301–304, 2016. DOI: 10.15199/48.2016.11.70