STUDIA Z AUTOMATYKI I INFORMATYKI
VOL. 41 – 2016

Piotr Kozierski[*], Talar Sadalla[†], Adam Dąbrowski[‡], Wojciech Giernacki[§]

ALLOPHONES IN AUTOMATIC SPEECH RECOGNITION

Keywords: Allophones, automatic speech recognition

1. Introduction

Automatic Speech Recognition (ASR) is increasingly used, and is applied in cars [17, 19], in so-called smart houses [10, 13], as well as in military applications [11]. ASR may be also helpful in communication for people after laryngectomy operation [15].

In the previous research [9], the usage of allophones in ASR for whispery speech has been discussed. It was concluded that the reduction of less frequently occurring allophones (by combining them with others) may have a positive effect on the obtained results. In this paper, it has been decided to extend previous research towards more complex approach and to use normal speech. During research more general data has been used (in [9] a very specific case has been used), and also four methods of allophones selection, for later combination with others, have been proposed.

In the second section, the speech recognition task and the programs used during research have been briefly presented. The third section contains the description of the allophones usage. In Section 4 there is information about the used speech corpus. The fifth section contains obtained results, which have been concluded in the last section.

2. Automatic speech recognition

The ASR task is to change the audio signal (which contains a speech) to the text, which is as close to the reference utterance, as possible. Therefore, based on the observation sequence $O_m = \{o_1, o_2, \ldots, o_m\}$ one should estimate the utterance $\hat{u}$, which has been recorded. Among all possible utterances u_i, the one, which maximizes function $p\left(u_i|O_m\right)$, is chosen [20]

$$\hat{u} = \arg\max_i p\left(u_i|O_m\right) = \arg\max_i \frac{p\left(O_m|u_i\right) \cdot p\left(u_i\right)}{p\left(O_m\right)} , \qquad (1)$$

[*] Poznan University of Technology, Faculty of Computing, Chair of Control and Systems Engineering, Division of Signal Processing and Electronic Systems, and also Poznan University of Technology, Faculty of Electrical Engineering, Institute of Control and Information Engineering, Piotrowo 3a street, 60-965 Poznań, e-mail: piotr.kozierski@gmail.com

[†] Poznan University of Technology, Faculty of Electrical Engineering, Institute of Control and Information Engineering, Piotrowo 3a street, 60-965 Poznań, e-mail: talar.h.sadalla@doctorate.put.poznan.pl

[‡] Poznan University of Technology, Faculty of Computing, Chair of Control and Systems Engineering, Division of Signal Processing and Electronic Systems, Piotrowo 3a street, 60-965 Poznań, e-mail: adam.dabrowski@put.poznan.pl

[§] Poznan University of Technology, Faculty of Electrical Engineering, Institute of Control and Information Engineering, Piotrowo 3a street, 60-965 Poznań, e-mail: wojciech.giernacki@put.poznan.pl

where $p(u_i)$ is the prior probability of the utterance u_i occurrence, $p(O_m)$ is the probability of all m observations, and $p(O_m|u_i)$ is the conditional probability that observation sequence occurs for the utterance u_i.

Each observation o_j corresponds to a signal frame (usually from 16 ms to 25 ms [12]), from which coefficients (e.g., Mel Frequency Cepstral Coefficients – MFCC) are obtained. Based on these coefficients, such a frame is associated with few, more probable phones (e.g., based on Gaussian Mixture Models – GMM – of phones, or triphones in more sophisticated cases). Taking into account few subsequent frames one can obtain many different connections of phones. These connections may be compared with lexicon (which contains the list of all possible phones connections, which form known words). With successive frames, the information about the possible word connections and their probabilities (from a language model) may be used.

The whole model, which contains the Acoustic Model (AM), lexicon and Language Model (LM), is constructed in the form of the Hidden Markov Model (HMM), where a state in the lowest level of this HMM model is associated with a single frame. Hence the estimated utterance can be considered as the sequence of m states $\hat{u} = S_m = \{s_1, s_2, \ldots, s_m\}$. The Viterbi algorithm is the method of finding this sequence [18].

The Kaldi toolkit [14], which is available under Apache v2.0 license, has been used during this research. All the implemented algorithms use Weighted Finite State Transducers (WFST), and therefore the OpenFST library [2] is necessary. Kaldi requires also other libraries and programs, such as Sequitur [3] to create and use the grapheme-to-phoneme (g2p) model, SRILM [16] for LM preparation, and also BLAS/LAPACK libraries for the linear algebra.

The standard MFCC parameters, with their first and second derivatives, have been used. The training path $\mathrm{mono} \rightarrow \mathrm{tri1} \rightarrow \mathrm{tri2a}$ (designations from Kaldi) has been chosen due to the relatively good quality of the speech recognition and the short training time (about 40 minutes using 3 CPUs) simultaneously.

3. Allophones in the used speech corpus

The basic components of speech are phonemes, which are typically used in ASR. Their number may be different depending on language, e.g., in Russian there are 55 phonemes, in American English there are 49 phonemes [6], whereas in Polish there are from 37 (SAMPA notation) to 39 (extended SAMPA notation) phonemes [7, 4]. However, each phoneme may be pronounced a little differently, depending on the context. Hence, within a single phoneme one can specify several different allophones. In Polish there are from 87 [21] to 91 [1] different allophones. Scripts have been prepared for all possibilities; however, in the whole corpus 11 allophones do not occur.

Additionally, in the IPA notation, which is available in Wiktionary (which was used to g2p creation), occur two symbols ”^{w}” and ”^{j}”, which mean labialization and palatalization, respectively. Hence, the whole number of phones, which have been taken into account, is 93 – all have been presented in [9], including the notation used in the Kaldi toolkit.

In the performed research, the usage of allophones instead of phonemes has been proposed by the authors. Unfortunately, some allophones are very rare, and thus the chance for a correct training of the HMM models for these phones is very low, and in some cases (e.g., when in the whole corpus allophone occurs only once) even impossible. Therefore, among the rarest allophones (number of occurrences in corpus less than 200; the most common allophone has 22,000 occurrences) possibilities of combining with other allophones of the same phoneme have been considered. Based on this, 40 possible combinations have been proposed, which have been shown in Tab. 1.

Tab. 1. The list of the less frequent allophones and the considered allophones changes

Allophone	Occurrences number (in whole corpus)	Considered combinations of allophones			
NI	1	NI→NY	NI→JN	NI→NOW	
OJ	1	OJ→O	OJ→OI	OJ→OO	
MI	2	MI→M	MI→MJ	MI→MY	
OI	16	OI→O	OI→OO		
NY	17	NY→N	NY→NJ	NY→NN	NY→NNJ
LY	23	LY→L	LY→LJ		
ZZJ	36	ZZJ→ZZ			
NJ	39	NJ→N	NJ→NN	NJ→NNJ	
DZZ	41				
AI	49	AI→A	AI→AA	AI→AJ	
MY	62	MY→M	MY→MJ		
TSJ	80	TSJ→TS			
JUP	89	JUP→I	JUP→II		
WW	90	WW→W	WW→WJ		
RY	97	RY→R	RY→RJ		
WJ	99	WJ→W			
WUP	99	WUP→U	WUP→UU		
NNJ	119	NNJ→N	NNJ→NN		
ZJ	128	ZJ→Z			
SSJ	181	SSJ→SS			

4. Speech corpus

The corpus used in this research contains over 9 hours of speech read by 33 different speakers and has been created by the authors of [8] – it is planned to share the corpus with other scientists in 2017 or 2018. The sentences come mainly from Andersen’s fairy tales, such as ”The Ugly Duckling”, ”The Fir Tree”, ”The Nightingale” and others.

Each speaker prepared recordings using his own device, therefore the recordings quality and the noise level are very diverse. All recordings have been saved in 48 kHz sampling frequency and in 16-bit sample depth.

5. Obtained results

During the selection of allophones which are to be changed into others, the results obtained from five different speakers (no. 8, 9, 13, 17 and 21) have been taken into account. The choice of these speakers was caused by the recordings length (a similar words number – from 3,000 to 4,400 words for each speaker; 18,635 words in total) and by the obtained results in general (not too good and not too weak). However, because of the relatively small size of the corpus, for each tested speaker the ASR model was trained separately, using all available data, except from this one tested speaker.

Four different strategies for the appointment of the set of allophones, which should be changed, have been considered by the authors – in every choice step one should look for:

a) the largest increase of the recognition quality,
b) the largest increase of the recognition quality to the number of converted allophones ratio,
c) (for cases in which the highest number of speakers have non-worsened results) the largest increase of the recognition quality,
d) (for cases in which the highest number of speakers have non-worsened results) the largest increase of the recognition quality to the number of converted allophones ratio.

Cases a-b and c-d are the same, however c-d strategies have been considered to robustify the choice in the cases, when for one speaker one can obtain considerable improvement of the speech recognition, but the decrease for the remaining four speakers.

The obtained results have been presented in Figure 1. Every step has been described by two lines. In the upper line, the considered change has been shown ("NI→NOW" means that all allophones NI have been changed into NOW allophones at the beginning, in both training and testing data, and then the normal ASR model training procedure has been performed). In the bottom line, 3 different pieces of information have been presented. The first value informs about the change in total number of errors ("-54" means that 54 errors less have been made in total after allophones changes in comparison to the previous step). The second number indicates the number of speakers for which the result has been worsened ("pos 3" means that for two speakers the speech recognition has been improved or has not changed in comparison to the previous step, and for 3 speakers the number of errors has been increased). The last value is the number of the allophone occurrence in the whole corpus.

In the Figure 1 one can see only the cases, for which the improvement of the speech recognition has been obtained (negative change in the errors number); however, simultaneously one can see all the cases, for which the improvement has been obtained. It means that after the allophones change MI→MJ there was no choice, because in the next step only the combination NY→N provided speech recognition improvement (reducing the errors number).

Therefore, based on the 4 strategies (a-d), three sets of the allophones changes have been found:

- case a) – MI→MJ, NY→N, NJ→NNJ;
- cases b) and d) – NI→NOW, TSJ→TS, MI→MY, WJ→W, ZZJ→ZZ, OJ→OO;
- case c) – NI→NOW, TSJ→TS, MI→MY, WJ→W, ZZJ→ZZ, ZJ→Z.

At the end, the speech recognition quality has been compared for these 3 sets of allophones, and also for the base case (all allophones, without any changes) and for phonemes-only approach. The quality of the speech recognition has been specified by the Word Error Rate (WER) index

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

where Subs is the number of substitutions, Del is the number of deletions, Ins is the number of insertions, and N_{utt} is the number of words in reference utterances. At this research step, 17 different speakers have been taken into account (48,659 words in sum). The results have been presented in Tab. 2.

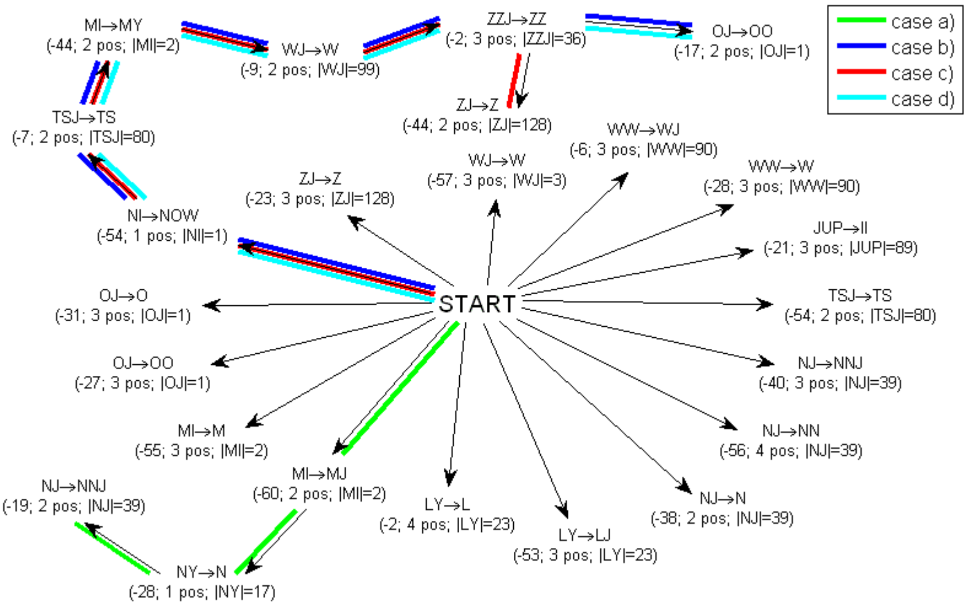

Fig. 1. The results obtained after allophones combination; information about changed allophones is the top row, and below – the information about the change of the total errors number, the number of speakers, for which the speech recognition is worse than before combination and the number of allophones in the whole corpus, respectively.

Tab. 2. Speech recognition quality based on the 17 different speakers

Case	Sum of errors Subs + Del + Ins	%WER
Base	8,195	16.842%
case a)	8,221	16.895%
cases b) and d)	8,143	16.735%
case c)	8,201	16.854%
phonemes approach	8,117	16.681%

6. Conclusions

The most general conclusion is that the allophones usage, when the normal speech is considered, does not make sense – in Tab. 2 one can see that phonemes approach provides the best speech recognition quality. However, it has been presented in the previous research that in the whispery speech the changing of few allophones into another provides better recognition quality, even in comparison to the phonemes-only approach [9].

It is surprising that in the chosen cases (a and c) the final results were worse than the base case (see Tab. 2), despite the fact that in every step only cases with better speech recognition were taken into account – this is due to the fact that the better recognition quality was obtained mainly for these 5 tested speakers, and for the remaining 12 speakers mainly decreases of the ASR quality have been noticed.

The value of %WER is not very good, but this is due to the relatively small amount of data used for training, as well as due to the quality of recordings prepared by the speakers. For comparison, in [5] (also for Polish) the phrase recognition rate was obtained between 83-92%.

After the detailed analysis one can say that past every of the considered changes (see Tab. 1) one is closer to the phonemes-only approach. Therefore, subsequent combinations should not have negative influence on the obtained results, but one can see that most of choices effects in deterioration of the ASR quality. Since the allophones combinations in many cases have a negative effect on the speech recognition quality, there must be some additional information in the IPA notation (which uses allophones).

Hence, the final conclusion is that the allophones usage makes sense, because there exists the potential to improve the speech recognition quality. However, the unlocking of this potential will be possible after the adaptation of ASR algorithms to the allophones approach. Therefore, this will be the focus of the future research.

REFERENCES

[1] Polish language - pronunciation - phones (in Polish). https://pl.wiktionary.org/wiki/Aneks:J%C4%99zyk_polski_-_wymowa_-_g%C5%82oski. access 11 I 2017.

[2] C. Allauzen, M. Riley, J. Schalkwyk, W. Skut, and M. Mohri. Openfst: A general and efficient eeighted finite-state transducer library. In *In Implementation and Application of Automata*, pages 11–23. Springer Berlin Heidelberg, 2007.

[3] M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, 2008.

[4] 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*, 7:79–97, 2003.

[5] P. Żelasko, B. Ziółko, T. Jadczyk, and D. Skurzok. Agh corpus of Polish speech. *Language Resources and Evaluation*, 50(3):585–601, 2015. DOI: 10.1007/s10579-015-9302-y.

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

[7] P. Kłosowski. Improving speech processing based on phonetics and phonology of Polish language. *Przegląd Elektrotechniczny*, 89(8):303–307, 2013.

[8] P. Kozierski, T. Sadalla, A. Dąbrowski, and D. Horla. Program for Polish whispery speech corpus creation (in Polish). *Informatyka, Automatyka, Pomiary w Gospodarce i Ochronie Środowiska (IAPGOŚ)*, accepted.

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

[10] Y. Mittal, P. Toshniwal, S. Sharma, D. Singhal, R. Gupta, and V. K. Mittal. A voice-controlled multi-functional smart home automation system. In *In 12-th IEEE India International Conference (INDICON)*, pages 1–6, New Delhi, December 2015.

[11] S. Pigeon, C. Swail, E. Geoffrois, G. Bruckner, D. V. Leeuwen, C. Teixeira, et al. *Use of Speech and Language Technology in Military Environments*. North Atlantic Treaty Organization, Montreal, Canada, 2005.

[12] B. Plannerer. *An Introduction to Speech Recognition.* Munich, Germany, March 2005.

[13] 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*, 17(1):127–144, 2013.

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

[15] 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*, 57(10):2448–2458, 2010.

[16] A. Stolcke. Srilm-an extensible language modeling toolkit. In *In Proc. Intl. Conf. Spoken Language Processing (INTERSPEECH)*, Denver, Colorado, September 2002.

[17] D. F. Syu, S. W. Syu, S. J. Ruan, Y. C. Huang, and C. K. Yang. Fpga implementation of automatic speech recognition system in a car environment. In *In 2015 IEEE 4th Global Conference on Consumer Electronics (GCCE)*, pages 485–486, October 2015.

[18] K. Szostek. Hmm models optimization and their use in speech recognition (in Polish). *Elektrotechnika i Elektronika*, 24(2):172–182, 2005.

[19] S. Wakisaka, K. Ishiwatari, K. Ito, T. Toge, and M. Tanaka. *Recognition Dictionary System Structure and Changeover Method of Speech Recognition System for Car Navigation.* Washington, 2000. U.S. Patent No. 6,112,174.

[20] S. Wydra. The use of mixed parameterization in Polish speech recognition system (in Polish). In *National Conference on Radiocommunication, Radiophony and Television (KKRRiT)*, Poznań, 2006.

[21] 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)*, pages 2601–2604, Barcelona, August 2003.

ABSTRACT

The common approach to speech recognition problem is the use of phonemes as basic parts of speech. The authors proposed allophones usage instead. For rarer allophones the conversion into other allophones (4 selection methods) has been proposed. Based on the obtained results one can say that the effective use of the additional information from the allophonic notation will not be possible without modification of currently used algorithms.

ALOFONY W AUTOMATYCZNYM ROZPOZNAWANIU MOWY

STRESZCZENIE

Typowym podejściem do zagadnienia rozpoznawania mowy jest branie pod uwagę fonemów, jako podstawowych części mowy. Zamiast tego autorzy zaproponowali wykorzystanie alofonów. Dla najrzadziej występujących alofonów zaproponowano ich zamianę na inne alofony – zaproponowano 4 metody wyboru głosek do zamiany. Na podstawie uzyskanych wyników stwierdzono, że efektywne wykorzystanie dodatkowych informacji, jakie niosą alofony, nie będzie możliwe bez modyfikacji obecnie dostępnych algorytmów.

Received: 2016-11-02
Accepted: 2017-01-04