HMM-based Breath and Filled Pauses Elimination in ASR

Piotr Żelasko[1], Tomasz Jadczyk[1,2] and Bartosz Ziółko[1,2]
[1]*Faculty of Computer Science, Electronics and Telecommunications, AGH University of Science and Technology, al. A. Mickiewicza 30, 30-059 Kraków, Poland*
www.dsp.agh.edu.pl
[2]*Techmo, Kraków, Poland*
techmo.pl
pzelasko@student.agh.edu.pl, {jadczyk, bziolko}@agh.edu.pl

Keywords: ASR, breath, breath detection, filled pause, filled pause detection, filler, filler detection, HMM, IVR, speech, speech recognition, spontaneous speech

Abstract: The phenomena of filled pauses and breaths pose a challenge to Automatic Speech Recognition (ASR) systems dealing with spontaneous speech, including recognizer modules in Interactive Voice Reponse (IVR) systems. We suggest a method based on Hidden Markov Models (HMM), which is easily integrated into HMM-based ASR systems and allows detection of those disturbances without incorporating additional parameters. Our method involves training the models of disturbances and their insertion in the phrase Markov chain between word-final and word-initial phoneme models. Application of the method in our ASR shows improvement of recognition results in Polish telephonic speech corpus LUNA.

1 INTRODUCTION

The Automatic Speech Recognition (ASR) has become a feature desired by many companies as a way to provide modern and ergonomic customer support via the telephone - it replaces the outdated Dual Tone Multi Frequency (DTMF) code navigation in Interactive Voice Response (IVR) systems dialogue menus, introducing more comfortable interface for their customers. In this scenario, the tasks of ASR systems are usually simple enough - recognize the word or a short phrase uttered by the end-user and provide recognition results to the overseeing IVR system. Commonly, the possibilities of dialogue options in such systems are limited, which leads to another simplification for the ASR - the dictionary is quite small, and may even vary from menu to menu to decrease its size even more.

One of the difficulties in such environments, however, is impossibility of guaranteeing, that the speaker will talk to the system as planned by the dialogue menu designers. A speaker rightfully assumes that they can speak naturally, sometimes producing utterances, phrases and sounds which do not exist in ASR systems dictionary, or cannot be classified as speech. This phenomenon is known as the out-of-vocabulary (OOV) utterance. A special case of OOV that interests us are sounds made by the user other than speech. We call them acoustic disturbances, and classify several of them: filled pauses (yyy, mmm, uhm, uh), breaths, impulsive noises (e.g. tapping the phone), coughing, blowing into the microphone, and similar. In normal physiological conditions, people breathe at about 12-20 breaths per minute (Konturek, 2007) (Ratan, 1993) and when they speak, their breath frequency decreases to about 10-12 breaths per minute (Igras and Ziółko, 2013a). As for the filled pauses, frequency of their usage depends largely on individual speakers – however, it can reach even up to 10 fillers per minute (Barczewska and Igras, 2012). Considering how frequent these events are, means have to be developed to handle them.

There exist some techniques for filled pause detection: cepstral variation based technique (Stouten and Martens, 2003), pitch-based technique (Goto et al., 1999) or formant-based technique (Audhkhasi et al., 2009), which work under assumption that the spectral variability of the filled pauses is low compared to speech (see figure 1). Breaths may be either actively looked for and detected (Igras and Ziółko, 2013b) or intentionally omitted - Boakye and Stolcke (Boakye and Stolcke, 2006) developed a speech activity detector based on Hidden Markov Models (HMM), which does not react to breath events. A typical breath spectrum may be seen on figure 2.

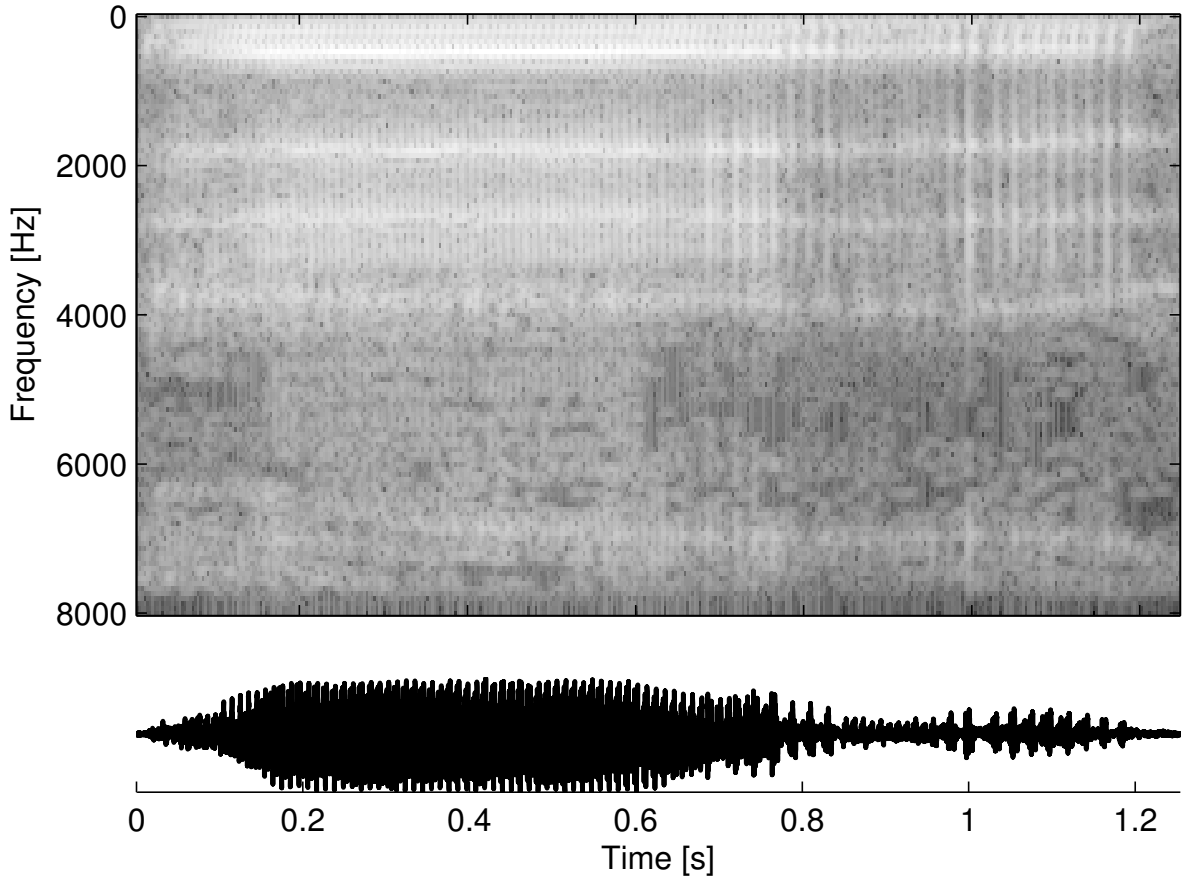

Figure 1: Time and spectral representation of an example of a filled pause. Notable features are its clearly distinguishable formants and little spectral variation over time.

We decided to investigate a different approach to the problem. Because our ASR (Ziółko et al., 2011) employs HMM to fit word models into segmented regions of speech, delivered by a voice activity detector (VAD), it often occurs that a breath or a filled pause is contained inside such a region. Those acoustic disturbances may occur not only on phrase boundaries, but also between words, if the speech is continuous. We seek a way to allow the recognizer module to deal with these disturbances without any additional parametrisation steps (such as fundamental frequency (F0) estimation or formants estimation), but based only on the actual set of parameters defined in our system. For the purpose of this research, we chose to use the 13 Mel Frequency Cepstral Coefficients (MFCC) along with their deltas and double deltas.

2 METHOD

We suggest a way to handle the speaker-generated acoustic disturbances, especially the breaths and filled pauses, by training their corresponding HMMs. Each acoustic disturbance model is then added to the phoneme model database of the ASR and is assigned to a group of models that we call the *sil* group. The idea behind introduction of group of models is to create a set of optional paths in the Markov chain of the word being detected, which allows skipping through these grouped models - or a part of them - if a better path can be found (i.e. the speaker had not uttered any of the modelled disturbances, but he could have).

In our experiment, the *sil* group consists of a silence model, a breath model and a filled pause model. We investigated two types of insertion of this group: in the first one, the possible transitions between disturbance models are fixed, allowing detection of only certain combinations of distrubances (Figure 3). For example, a common sequence of *(breath, filled pause)* would be detected, but a less common one: *(filled pause, breath)* is not modelled by this type of insertion. The second variant of insertion has no fixed transitions – every possible sequence of disturbances can be modelled.

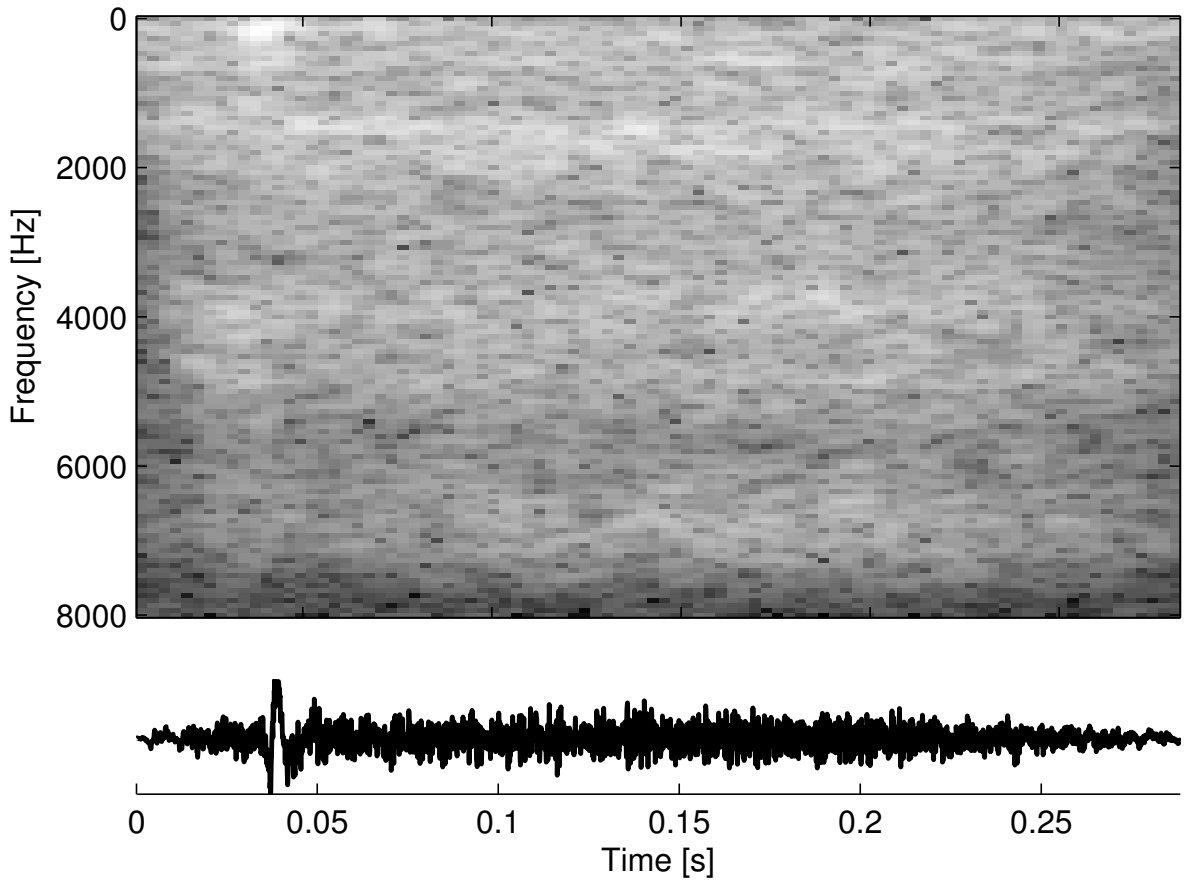

Figure 2: Time and spectral representation of an example of a breath. Unlike the filled pause, its spectrum is similiar to the noise, although some indication of formants is observed.

The models are inserted before the beginning of each word and after its end, and in case of a phrase they are inserted between each word and on the phrase boundaries. An example of this insertion - in our dictionary, there is a Polish phrase *o tak* (in English: *oh yes*), which consists of four phoneme models {*O, T, A, K*} and three optional silence models {*SIL*}, represented by a Markov chain {*SIL, O, SIL, T, A, K, SIL*}. After addition of acoustic disturbances models {*BREATH, FILLER*}, the exact Markov chain for this word looks like the following: {*SIL, BREATH, FILLER, O, SIL, BREATH, FILLER, T, A, K, SIL, BREATH, FILLER*}, where every non-phoneme model may be omitted. Additionally, we also checked the scenario which does not involve insertion of silence models between words in a phrase, leaving them only on phrases beginning and end.

In order to be able to control the sensitivity of disturbances detection, we implemented an option to include transition probability penalty for the disturbance models. The idea is to multiply between-model transitions probabilities by weighting factors, which sum to unity. The obligatory model transition weight (i.e. the transition from the last phoneme of a preceding word to the first phoneme of a following word) is set to a certain value, and the difference between unity and this value is divided equally by the number of optional (disturbance) models inserted after word-final phoneme model and assigned to each of them as their transition probability weight. We investigated a total of three scenarios: no penalty included (equal transition probability weighting factors), 50-50 penalty (0.5 weighting factor for the next phone model, and 0.5 to split between disturbance models) and 75-25 penalty (0.75 weighting factor for the next phone model, and 0.25 to split between disturbance models). For example, on Figure 3, if a 50-50 penalty was applied, the weighting factors would be 0.5 for O into T transition and 0.(6) for O into SIL, O into BREATH and O into FILLER transitions.

In order to perform the training of the acoustic disturbances models, we prepared manual annotations of breaths and filled pauses. Our data consists of recordings of Polish translations from Europarliament ses-

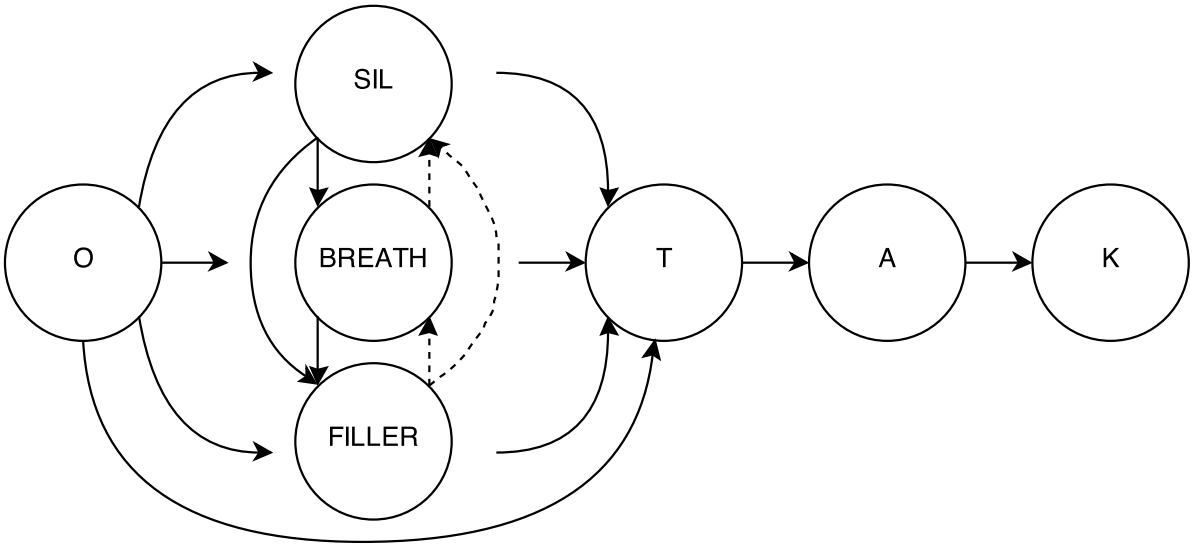

Figure 3: Example of disturbance models (SIL - silence model, BREATH - breath model, FILLER - filled pause model) insertion into a Markov chain of a Polish phrase *o tak* (English equivalent: *oh yes*). If the transitions marked by the dashed arrows are removed, then direction of transitions between disturbance models is fixed, disallowing a previous disturbance model to appear again in the alignment. Otherwise, the transition direction is arbitrary - any combinations of disturbances is being modelled.

sions (Gollan et al., 2005) – 30 minutes, Polish radio auditions – 60 minutes and LUNA, a corpus of Polish spontaneous telephonic speech (Marciniak, 2010) – we annotated breaths and fillers in 80 minutes of this corpus. A summary containing statistics of our data is presented in Table 1. Each disturbance model was trained to consist of three HMM states, with state emission probability computed with use of a three-component Gaussian Mixture Model (GMM).

Table 1: Statistics of training data used in acoustic disturbances model creation.

Disturbance	Breath	Filled Pause
Mean duration [ms]	372	417
Minimal duration [ms]	101	80
Maximal duration [ms]	1208	1350
Total duration [s]	217	249
Number of occurences	584	597

3 RESULTS

The testing corpus were another 80 minutes of LUNA, which is 74 recordings, however, they were different ones than used in the training. Each recording from LUNA has an annotation of phrases being spoken (in this context, a phrase is a sequence of words separated by no more than 100ms silent pause), but does not contain information about appearances of breaths or filled pauses. We established, based on the training data from LUNA (which contains 1258 phrases, that is 12% of all phrases in LUNA) that breaths appear in about 13% of phrases and filled pauses appear in about 15% of phrases.

We compared the recognition results of our ASR in several different scenarios. In the first one, we present three variants of configuration: NM – no additional models inserted between words in phrases, SM – silence models inserted between words in phrases and SM+DM, where both silence models and disturbance models were inserted between the words. The results are presented in Table 2. In these tests, the transitions between disturbance models were fixed, as presented on Figure 3. Later, penalties for transitioning into disturbance model were applied to check for improvement. 50-50 and 75-25 denote different types of penalties, as explained in the section 2; NP stands for no penalty.

Table 2: Results of correct recognition percentage in different data sets in three variants: NM (no models inserted between words), SM (silence models inserted) and SM+DM (silence and disturbances models inserted). The directions of transitions between disturbance models are fixed.

	NP	50-50	75-25
NM	80.3%	–	–
SM	77.6%	–	–
SM+DM	78.3%	79.2%	79.1%

These results suggested a problem with false

alarms raising at the current sensitivity level of disturbance (and silence) detection. Therefore, in the next phase of testing, we continued to investigate the possibility of calibrating the system by forcing the transitions from phoneme models to disturbance and silence models to be less probable with help of transition penalties. We also noticed that a major factor affecting negatively the recognition might be the insertion of silence model, so we checked the results of recognition without inserting it between the words. We also introduced the arbitrary transition direction between the disturbance and silence models, in order to be able to detect every combination of disturbances. The results are shown in Table 3.

Table 3: Results of correct recognition percentage in different data sets in two variants: with only silence models inserted between words (SM) and with silence and disturbances models inserted between words (SM+DM). The directions of transitions between disturbance models are arbitrary.

	NP	50-50	75-25
SM+DM	80.2%	80.1%	80%
DM	81.1%	81.1%	81%

Introduction of the arbitrary transition directions between disturbance models improved the effectiveness of disturbances detection from 78.3% to 80.2%. Further improvement was achieved by removing the silence model from between the words at 81.1%, which is higher score than NM variant at 80.3%. The penalty system is unsatisfactory: it offers no improvement with introduction of arbitrary model transitions.

In order to verify whether the best achieved outcome (i. e. recognition rate of 81.1%) is a significant improvement compared to what we already had without including disturbances models (i. e. recognition rate of 80.3%), we took the 1774 recognition results from both bare system and enhanced system and performed bootstrapping on them (Bisani and Ney, 2004), which resulted in 1000 bootstrap populations of recognition results for both system variants. Because both variants of the system were tested on the same data, we made sure that the bootstrap populations were also the same, allowing us to directly compare recognition rate of each system variant in every population. We then calculated the mean recognition rate, its standard deviation and confidence intervals for both systems (tab. 4). Histograms of this data are presented on figure 4.

As the last step, we calculated the *probability of improvement* (*poi*) measure, defined by

$$poi = \frac{1}{B}\sum_{b=1}^{B}\Theta(\Delta RR_b)\,, \qquad (1)$$

Table 4: Mean, standard deviation and 90% confidence intervals of recognition rate for two system variants - system with no modifications and system with disturbance models, arbitrary transition directions, no penalties and no silence models. We also present statistics for differences between recognition rate in each bootstrap sample in the *Improv.* row.

Rec. rate	Mean	St. dev.	Conf. interv.
Unmod. sys.	80.3%	0.92%	(78.7 - 81.8)%
Enh. sys.	81.1%	0.92%	(79.5 - 82.6)%
Improv.	0.8%	0.52%	(-0.1 - 1.7)%

where B is the number of bootstrap samples, $\Theta()$ is the Heaviside function and ΔRR_b is the recognition rate difference between enhanced system and unmodified system for b'th bootstrap sample. This measure shows the percentage of bootstrap samples in which recognition rate has been improved in the enhanced system. In our case, *poi* amounted to 93.95%, leading to conclusion that system is improved by introduction of disturbance models. More details on this measure may be found in (Bisani and Ney, 2004).

4 CONCLUSIONS

Our approach of handling acoustic disturbances, such as breaths or filled pauses, improves the resuts of spontaneous telephonic speech recognition. In order to achieve the improvement, arbitrary transitions between disturbance models are preferred and insertion of silence models inside a phrase is discouraged. The method is useful in ASR systems, which incorporate Hidden Markov Models in the recognition task, serving as an extension to the existing system. Application of the method can be found in any scenario, where ASR system has to deal with spontaneous speech, an example of such scenario being recognition of user commands in IVR system menu.

ACKNOWLEDGEMENTS

This work was supported by LIDER/37/69/L-3/11/ NCBR/2012 grant. We thank Magdalena Igras for assistance in gathering recordings and transcriptions of filled pauses and breaths and Jakub Gałka for advices regarding statistical analysis of our results.

Narodowe Centrum Badań i Rozwoju

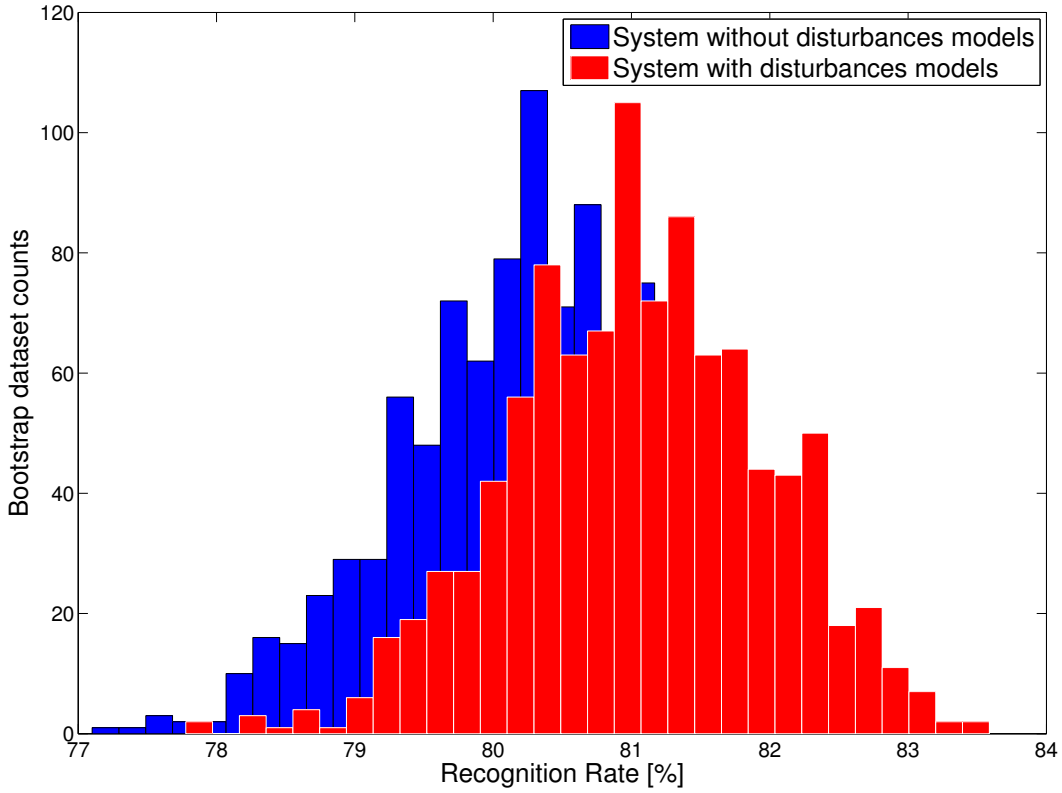

Figure 4: Histograms of bootstrapped recognition results from umodified (blue) and enhanced (red) variants of systems.

REFERENCES

Audhkhasi, K., Kandhway, K., Deshmukh, O., and Verma, A. (2009). Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4857–4860.

Barczewska, K. and Igras, M. (2012). Detection of disfluencies in speech signal. In *Young scientists towards the challenges of modern technology: 7th international PhD students and young scientists conference in Warsaw*.

Bisani, M. and Ney, H. (2004). Bootstrap estimates for confidence intervals in asr performance evaluation. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, volume 1, pages I–409–12 vol.1.

Boakye, K. and Stolcke, A. (2006). Improved speech activity detection using cross-channel features for recognition of multiparty meetings. In *Proc. of INTERSPEECH*, pages 1962–1965.

Gollan, C., Bisani, M., Kanthak, S., Schluter, R., and Ney, H. (2005). Cross domain automatic transcription on the tc-star epps corpus. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 825–828.

Goto, M., Itou, K., and Hayamizu, S. (1999). A real-time filled pause detection system for spontaneous speech recognition. In *Proc. of Eurospeech*, pages 227–230.

Igras, M. and Ziółko, B. (2013a). Modelowanie i detekcja oddechu w sygnale akustycznym. In *Proc. of Modelowanie i Pomiary w Medycynie*.

Igras, M. and Ziółko, B. (2013b). Wavelet method for breath detection in audio signals. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6.

Konturek, S. (2007). *Fizjologia człowieka. Podrecznik dla studentów medycyny*. Elsevier Urban & Partner.

Marciniak, M., editor (2010). *Anotowany korpus dialogów telefonicznych*. Akademicka Oficyna Wydawnicza EXIT, Warsaw.

Ratan, V. (1993). *Handbook of Human Physiology*. Jaypee.

Stouten, F. and Martens, J. (2003). A feature-based filled pause detection technique for dutch. In *IEEE Intl Workshop on ASRU*, pages 309–314.

Ziółko, M., Gałka, J., Ziółko, B., Jadczyk, T., Skurzok, D., and Mąsior, M. (2011). Automatic speech recognition system dedicated for Polish. *Proceedings of Interspeech, Florence*.