Bimodal Speech Recognition Fusing Audio-Visual Modalities

Alexey Karpov[1(✉)], Alexander Ronzhin[1], Irina Kipyatkova[1],
Andrey Ronzhin[1], Vasilisa Verkhodanova[1], Anton Saveliev[1],
and Milos Zelezny[2]

[1] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, SPIIRAS, Saint-Petersburg, Russian Federation, Russia
{karpov, ronzhinal, kipyatkova, ronzhin}@iias.spb.su
[2] University of West Bohemia, Pilsen, Czech Republic
zelezny@kky.zcu.cz
http://hci.nw.ru
http://www.kky.zcu.cz

Abstract. In this paper, we present a novel bimodal speech recognition technique that fuses both audio information (sound signal) and visual information (movements of lips) for Russian speech recognition. We propose an architecture of the automatic system for bimodal recognition of audio-visual speech, which uses one stationary microphone Oktava and one high-speed camera JAI Pulnix (200 frames per second at 640 × 480 pixels) to get audio and video signals. We describe also developed software for audio-visual speech database recording, phonemic and visemic structures of the Russian language, as well as probabilistic models of bimodal speech units based on Coupled Hidden Markov Models. Realization of a transformation method from a Coupled Hidden Markov Model into an equivalent 2-stream Hidden Markov Model is presented as well.

Keywords: Automatic speech recognition · Audio-Visual speech processing · Speech technology · Information fusion · Automatic Lip-reading · Bimodal system and interface

1 Introduction

At present, there are some Russian automatic speech recognition (ASR) systems based on audio-only signal captured via a microphone developed by some International industrial companies, such as Google (Google Now software), Nuance (Dragon Naturally Speaking software), Apple (Siri software for iPhone 6), Microsoft (software for Xbox One and Cortana for Windows 8), Samsung, as well as by some Russian commercial companies, leading state Universities and Institutes of the Russian Academy of Sciences [1, 2].

However, present performance of audio-only Russian ASR is not satisfactory for the most of end-users. However, it is well known from some recent works [3–6], that audio and visual modalities (cues) of speech supplement each other very well and their joint multimodal processing can improve both accuracy and robustness of ASR.

M. Kurosu (Ed.): HCI 2016, Part II, LNCS 9732, pp. 170–179, 2016.
DOI: 10.1007/978-3-319-39516-6_16

In this paper, we present a bimodal automatic speech recognition technique that fuses both audio information (speech sound signal) and visual information (lip movements) for the Russian speech recognition.

The paper is structured as follows: Sect. 2 presents general architecture of the bimodal speech recognition system, Sect. 3 describes probabilistic models of audio-visual speech units based on Coupled Hidden Markov Models, and conclusions are outlined in Sect. 4.

2 Architecture of the Bimodal Speech Recognition System

General architecture of a typical bimodal speech recognition system, which fuses both audio and visual modalities, is presented in Fig. 1. As any state-of-the-art speech recognition system our system operates in two modes: (1) model training and (2) speech decoding/recognition.

In this system, we use one high-speed camera JAI Pulnix (200 fps at 640 × 480 pixels) and one dynamic microphone Oktava in order to capture both video and audio signals. High frequency of video frames is crucial for analysis of dynamical images, since visible articulation organs (lips, teeth, tip of tongue) change their configuration quite fast at speech production and duration of some phonemes (e.g. explosive consonants) is within 10–20 ms (duration of each video at 25 fps frame is 40 ms that is too long). So recordings made by a standard camera with 25–30 fps cannot catch fast dynamics of lips movements and most of the important information is missing in these signals.

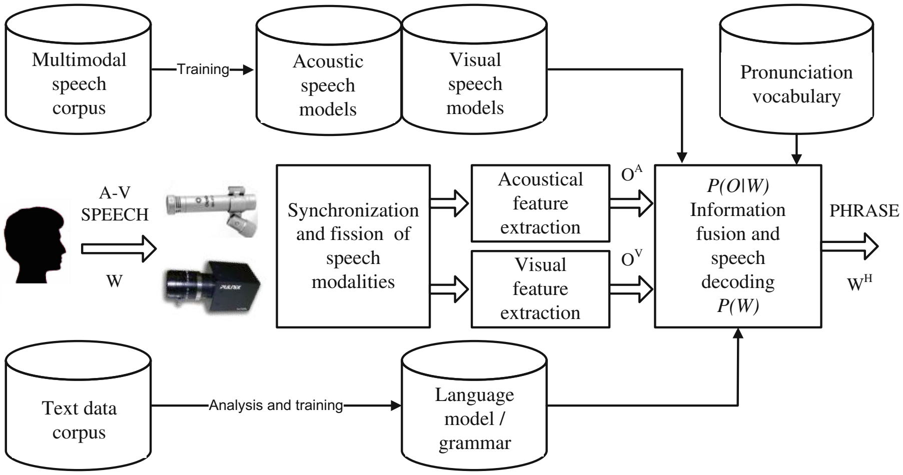

Fig. 1. General architecture of the bimodal speech recognition system

In our bimodal speech recognition system, 12-dimentional Mel-frequency cepstral coefficients (MFCC) are extracted as acoustical features, calculated from 26 channel filter bank analysis of 20 ms long frames with 10 ms step [7]. Visual features of speech are calculated as a result of the following signal processing steps using the open source computer vision library OpenCV [8]: multi-scale face detection in video frames from a video-camera using a cascade classifier with AdaBoost method based on the Viola-Jones algorithm [9, 10] with a trained face model; mouth region detection with two cascade classifiers (for a mouth and mouth-with-beard) within the lower part of the face [11]; normalization of detected mouth image region to 32 × 32 pixels; mapping to a 32-dimentional feature vector using the principal component analysis (PCA); visual feature mean normalization; viseme-based linear discriminant analysis (LDA). The video signal processing module produces 10-dimentional articulatory feature vectors.

In order to train probabilistic language models and acoustic-visemic models, text and audio-visual speech corpora are required. Figure 2 shows the architecture of software for audio-visual speech database recording. The audio-visual speech recording system is intended for formation of an audio-visual speech database to train probabilistic models of the speech recognition system.

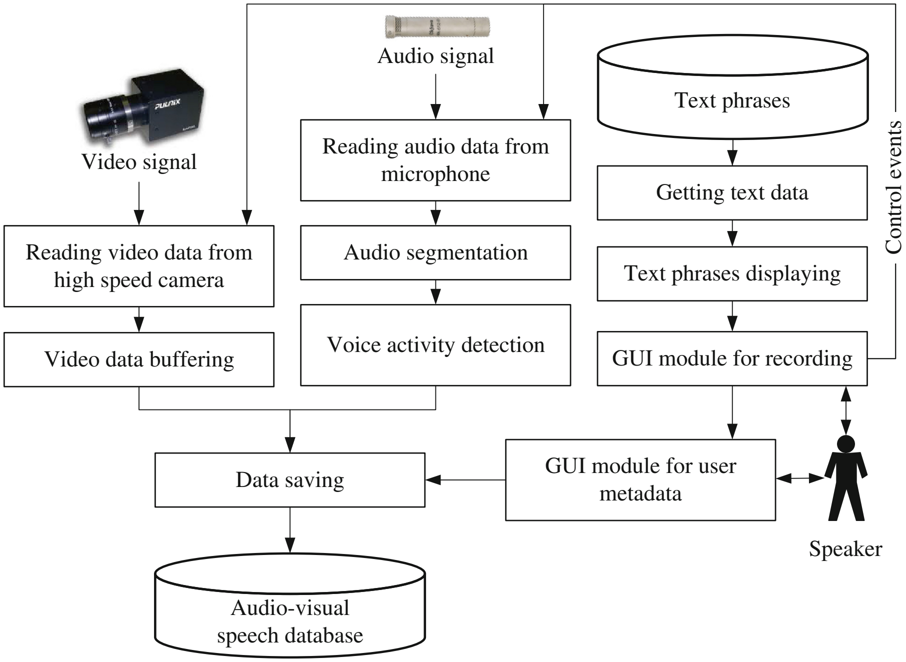

Fig. 2. Architecture of the audio-visual speech database recording software

The software complex consists of four main modules: (1) video data capturing and buffering; (2) audio data capturing, processing and segmentation; (3) text data displaying; (4) GUI for interaction with a user (speaker). The developed software has two GUI modules for interaction with user and receiving metadata. The modules of GUI provide two modes for data recording: (1) an “expert” mode, where the user manages start and end points; (2) an “automatic” mode, where fragments of recording are determined by a voice activity detection (VAD) method. In the “automatic” mode, capturing and processing of audio signal is carrying out continuously, as well as video data is buffered in RAM memory for the last 60 frames (300 ms at 200 fps). This buffering option is based on some aspects of human speech production. After the recording phrase, audio and video data of a current speaker are saved into the speech database; for synchronizing audio and video signals, the software calculates frame mistiming.

One major problem in machine audio-visual speech recognition is to implement a correct method for synchronization and unification of different speech modalities. The problem is that the two modalities naturally become desynchronized, i.e., streams of corresponding phonemes and visemes are not perfectly synchronous in real life due to natural constraints in the human speech production process, inertia in human articulation organs, and coarticulation (interdependence and interaction of adjacent elements in spoken speech) which has different effects on acoustic and visual speech components, leading to desynchronization. It is known that visemes always lead in phone-viseme pairs; at that in the beginning of a phrase visual speech units usually lead more noticeably (up to 150-200 ms for stressed rounded vowels) over the corresponding phonemes than in the central or ending part of the phrase.

3 Probabilistic Models of Audio-Visual Speech Units

To take into account the temporal desynchronization between the streams of corresponding acoustic and visual features, which is natural for human speech, we propose to apply Coupled Hidden Markov Models [3, 12]. Figure 3 presents a topology of such a model for an audio-visual speech unit (a phoneme–viseme pair) with several states for each stream of feature vectors. Circles denote HMM states that are hidden for observation; squares indicate mixtures of normal distributions of observation vectors in the states. A coupled Hidden markov model (CHMM) is a set of parallel HMMs, one per information flow (modality). Model states at some moment of time t for each HMM depend on hidden states at time moment *t-1* for all parallel HMMs. Thus, the entire state of a CHMM is defined by the collection of states for two parallel HMMs. One advantage of such a topology is that it lets several streams of feature vectors independently walk through the model states, which gives us a possibility to model admissible temporal inconsistencies in both audio and video data.

In the topology of CHMM audio-visual speech units we use three hidden states per each parallel stream of feature vectors, assuming that the first states correspond to a dynamical transition from the previous speech unit, the third states, to a transition to the next unit, and the second states of the unified model (the most lengthy ones) correspond to the stationary central segment of the speech unit. In order to define a CHMM $\lambda = \ <L, D, B, \gamma>$ for an audio-visual speech unit, we have to specify the following parameters:

(1) Number of hidden states in the model– L (states for speech audio and video modalities are shown with circles in Fig. 3 and denoted as A and V respectively).

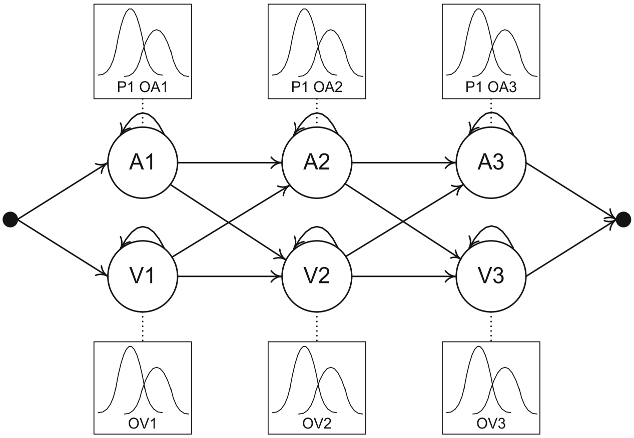

Fig. 3. Topology of a Coupled Hidden Markov Model for an audio-visual unit

(2) Matrix of transition probabilities between model states– $D = \{d_{ij}\}$, $1 \leq i \leq L, 1 \leq j \leq L$

(3) Probability distributions for the feature vector in model states (shown in squares in Fig. 3) – $B = \{b_j(O)\}$. We use mixtures of Gaussian distributions:

$$b_j(O) = \sum_{m=1}^{M} c_{jm} N(O,\ \mu_{jm},\ \sigma_{jm}^2), \sum_{m=1}^{M} c_{jm} = 1, 1 \leq j \leq L,$$

where O is the feature vector being modeled (for the audio or video signal), C_{jm} is the weight coefficient of component m in a state j, N is the distribution density (usually a Gaussian distribution density) with mean (expectation) μ_{jm} and variance (standard deviation) σ_{jm}^2 for mixture component m in the state j, M is the number of Gaussian components in the mixture (up to 16).

(4) Information weights (importance) $\gamma = \{\gamma^A,\ \gamma^V\}$ of speech modalities (audio and video streams); they are tuned during system training or adaptation, and they always sum up to the constant: $\gamma^A + \gamma^V = 2$.

The Russian language contains several dozens of various context-independent phonemes (different researchers and phoneticians distinguish 40–50 phonemes, in our system we use 48 phonemes, see Table 1), so there are as many different CHMMs in the automatic speech recognition system. However, there exist fewer number of various visemes in Russian speech; only about 10–12 depending on the speaker's articulation (in our system we use 10 visemes, see Table 1) [13, 14]. Each CHMM represents one phoneme-viseme pair, and to model audio speech signals, we need more HMMs than for the visual speech modeling only.

In order to cope with this problem, it is proposed to tie distributions of observation vectors for visual components in the states of different CHMMs. The total number of CHMMs in the system equals the number of phonemes being recognized, but some models have common states and parameters; it simplifies the training process. Figure 4 shows the scheme of tying parameters of a CHMMs pair for 2 phonemes corresponding to 1 viseme. After tying the output densities of corresponding viseme models according to the mapping in Fig. 4, we can get 30 tied probability densities for the visual data stream and 144 untied ones for the acoustical feature stream.

A simple method to transform a CHMM into an equivalent HMM, which keeps all the properties of the former model, was proposed in [15] and later used by the authors in [16]. This method is used in our recognition system as well. Transformed HMM for an

Table 1. Visemes and phoneme-to-viseme mapping for Russian speech

Viseme	Viseme class	Corresponding phonemes
/v1/	silence (neutral position)	/sil/(pause)
/v2/	wide-opened mouth unrounded vowels	/a/,/a!/,/e!/
/v3/	unrounded vowels (unstressed and stressed)	/i/,/i!/,/y/,/y!/,/e/
/v4/	rounded vowels	/o!/,/u/,/u!/
/v5/	labial consonants (hard and soft)	/b/,/b’/,/p/,/p’/,/m/,/m’/
/v6/	labio-dental consonants	/f/,/f’/,/v/,/v’/
/v7/	alveolar fricatives	/sh/,/zh/,/ch/,/sch/
/v8/	alveolar sonorants	/l/,/l’/,/r/,/r’/
/v9/	dental consonants	/d/,/d’/,/t/,/t’/,/n/,/n’/, /s/,/s’/,/z/,/z’/,/c/
/v10/	velar consonants	/g/,/g’/,/k/,/k’/,/h/,/h’/,/j/

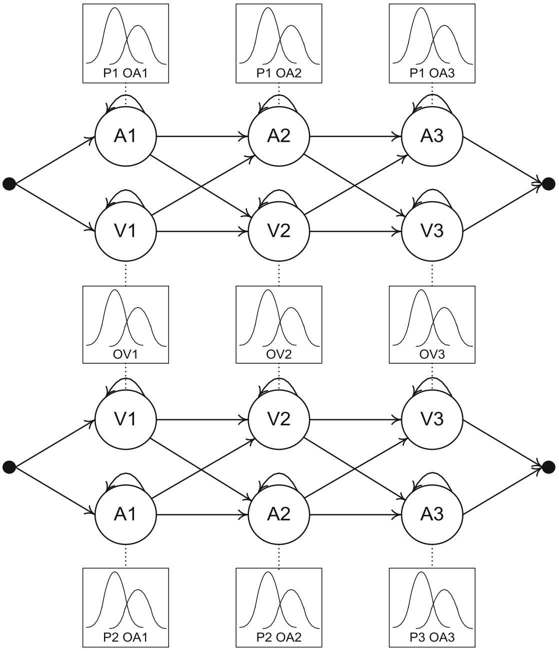

Fig. 4. Tying probability distributions of a pair of the Coupled Hidden Markov Models for two phonemes corresponding to one viseme.

audio-visual speech unit contains all the combinations of parallel states of the corresponding CHMM. In the CHMM model, the two streams are independent; the output distribution of a joint state is calculated by the output densities of both streams. In the equivalent 2-stream HMM, the output distribution is obtained as a product of the two output densities. To avoid tripling of the output densities in the model, it is proposed to tie the appropriate output densities in the 2-stream HMM according to CHMM-to-HMM conversion of the hidden states. The resulting 2-stream HMM is shown in Fig. 5. We use CHMM with 3 emitting states per feature stream. Therefore, all their combinations produce 9 states in the equivalent left-to-right HMM. Increment of the number of the states in comparison with the original CHMM increases the memory allocated for the model, but does not reduce the speed of speech decoding. The parameters of 2-stream HMMs are obtained by the Baum-Welch (expectation-maximization) algorithm with maximum likelihood estimation using bimodal training data.

In order to recognize (decode) speech, we apply a modified token passing algorithm based on the Viterbi optimization algorithm for multi-threaded HMMs [7, 17], which finds probabilities of generating observation symbols (sequences of feature vectors) in this model and sequences of the model's hidden states traversed along the way. The essence of our method is as follows: to model possible phrases, we construct a unified probabilistic model (graph) with all possible transitions between HMMs of minimal speech units (constrained by the recognition dictionary) and between HMMs of individual words (constrained by the language model). Then with dynamical programming (Viterbi) we find the maximum likelihood sequence/path of hidden states (that contain information about speech units) in the model to generate the processed sequence of observations [7]. As a result of decoding the speech signal, the automated system can output one or several best recognition hypotheses of the said phrase. The final solution for choosing the recognition hypothesis is made by maximizing the probabilities of generating hypotheses obtained in signal analysis.

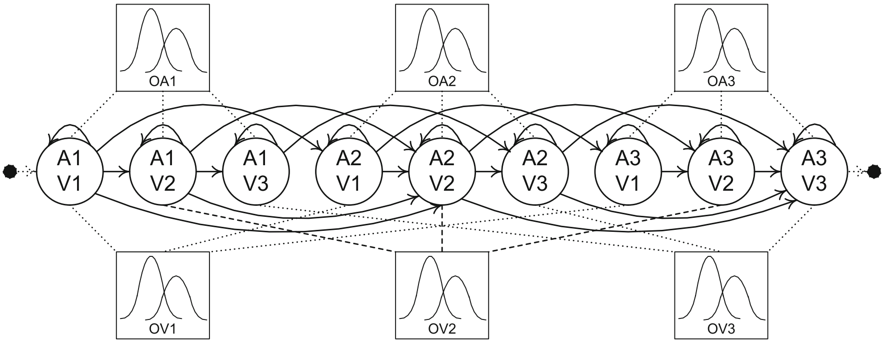

Fig. 5. Transformation of a CHMM into the equivalent 2-stream HMM

The AVSR system processes acoustical and visual observations in parallel, and it has to weight the informativity of one speech modality over the other. In standard CHMMs, it is made by setting audio and video stream weights and using them as exponents of the observation probabilities. However, we suppose that some phoneme and viseme models may be more reliable than others in varying environment and their contribution to the overall recognition performance may be bigger. So we propose to assign individual modality significance weights to each phoneme-viseme model. In this case, the observation probabilities in hidden states of HMMs are calculated as:

$$P(O_t|\lambda_{avunit}) = \prod_{s\in\{A,V\}} P(O_t^s|\lambda_{avunit})^{\gamma_{avunit}^s}$$

where O_t is the audio-visual observation vector at time *t,* whereas O_t^s represents the observation vector of one (audio or visual) stream *s* at time *t,* λ_{avunit} represents HMM parameters of a particular viseme-phoneme model, and γ_{avunit}^s means the significance weight of visual/audio stream for the given AV speech unit model.

For experimental research, we are collecting an audio-visual Russian speech database. According to our preliminary results the bimodal speech recognition that fuses both audio and visual modalities allows increasing the accuracy (word recognition rate) with respect to unimodal audio- or video-based speech recognition systems, especially in noisy conditions. In our future research, the developed multimodal ASR system will be a part of the universal assistive information technology [18, 19].

4 Conclusion

In the paper, we have presented the new bimodal speech recognition technique that fuses audio and visual information for the Russian speech recognition. We have proposed the general architecture of the automatic system for bimodal recognition of audio-visual speech; it uses the stationary microphone Oktava and the high-speed camera JAI Pulnix to get audio and video signals. We have also described the architecture of the developed software for audio-visual speech database recording, phonemic and visemic structures of the Russian language, as well as the probabilistic models of audio-visual speech units based on Coupled Hidden Markov Models and equivalent 2-stream Hidden Markov Models. For the experimental research, we have collected a part of the audio-visual Russian speech database. According to our preliminary results the bimodal speech recognition that fuses both audio and visual modalities allows improving the word recognition rate with respect to unimodal audio- or video-based speech recognition systems, especially in noisy conditions.

Acknowledgements. This research is financially supported by the Ministry of Education and Science of the Russian Federation, agreement № 14.616.21.0056 (reference RFMEFI61615X0056), project “Research and development of audio-visual speech recognition system based on a microphone and a high-speed camera”, as well as by the Czech Ministry of Education, Youth and Sports, project № LO1506.

References

1. Karpov, A., Markov, K., Kipyatkova, I., Vazhenina, D., Ronzhin, A.: Large vocabulary Russian speech recognition using syntactico-statistical language modeling. Speech Commun. **56**(1), 213–228 (2014)
2. Besacier, L., Barnard, E., Karpov, A., Schultz, T.: Automatic speech recognition for under-resourced languages: A survey. Speech Commun. **56**(1), 85–100 (2014)
3. Katsaggelos, A., Bahaadini, S., Molina, R.: Audio-visual fusion: challenges and new technologies. Proc. IEEE **103**(9), 1635–1653 (2015)
4. Stewart, D., Seymour, R., Pass, A., Ming, J.: Robust audio-visual speech recognition under noisy audio-video conditions. IEEE Trans. Cybern. **44**(2), 175–184 (2014)
5. Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H., Ogata, T.: Audio-visual speech recognition using deep learning. Appl. Intell. **42**, 722–737 (2015)
6. Deng, L., Li, X.: Machine learning paradigms for speech recognition: an overview. IEEE Transa. Audio Speech Lang. Process. **21**(5), 1060–1089 (2013)
7. Young, S., et al.: The HTK Book (for HTK Version 3.4). Cambridge University Press, Cambridge (2006)
8. Kaehler, A., Bradsky, G.: Learning OpenCV 3. O'Reilly Media, California (2015)
9. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition CVPR-2001, USA, pp. 511–518 (2001)
10. Liang, L., Liu, X., Zhao, Y., Pi, X., Nefian, A.: Speaker independent audio-visual continuous speech recognition. In: Proceedings of the International Conferenceon Multimedia and Expo ICME 2002, Lausanne, Switzerland, pp. 25–28 (2002)
11. Castrillyn, M., Deniz, O., Hernandez, D., Lorenzo, J.: A comparison of face and facial feature detectors based on the Viola-Jones general object detection framework. Mach. Vis. Appl. **22**(3), 481–494 (2011)
12. Nefian, A.V., Liang, L.H., Pi, X., Xiaoxiang, X., Mao, C., Murphy, K.: A coupled HMM for audio-visual speech recognition. In: Proceedings of the International Conference ICASSP-2002, Orlando, USA, pp. 2013–2016 (2002)
13. Karpov, A.A.: An automatic multimodal speech recognition system with audio and video information. Autom. Remote Control **75**(12), 2190–2200 (2014)
14. Karpov, A., Kipyatkova, I., Železný, M.: A framework for recording audio-visual speech corpora with a microphone and a high-speed camera. In: Ronzhin, A., Potapova, R., Delic, V. (eds.) SPECOM 2014. LNCS, vol. 8773, pp. 50–57. Springer, Heidelberg (2014)
15. Chu, S.M., Huang, T.S.: Multi-modal sensory fusion with application to audio-visual speech recognition. In: Proceedings of the Multi-Modal Speech Recognition Workshop-2002, Greensboro, USA (2002)
16. Karpov, A., Ronzhin, A., Markov, K., Zelezny, M.: Viseme-dependent weight optimization for CHMM-based audio-visual speech recognition. In: Proceedings of the International Conference, INTERSPEECH-2010, ISCA Association, Makuhari, Japan, pp. 2678–2681 (2010)
17. Benesty, J., Sondhi, M., Huang, Y., et al.: Springer Handbook of Speech Processing. Springer, New York (2008)

18. Karpov, A., Ronzhin, A.: A universal assistive technology with multimodal input and multimedia output interfaces. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2014, Part I. LNCS, vol. 8513, pp. 369–378. Springer, Heidelberg (2014)
19. Karpov, A., Ronzhin, A., Kipyatkova, I.: Automatic analysis of speech and acoustic events for ambient assisted living. In: Antona, M., Stephanidis, C. (eds.) UAHCI 2015. LNCS, vol. 9176, pp. 455–463. Springer, Heidelberg (2015)