Computer Science • 19(1) 2018 https://doi.org/10.7494/csci.2018.19.1.2398

Tomasz Jadczyk

AUDIO-VISUAL SPEECH-PROCESSING SYSTEM FOR POLISH APPLICABLE TO HUMAN-COMPUTER INTERACTION

Abstract *This paper describes an audio-visual speech recognition system for the Polish language as well as a set of performance tests under various acoustic conditions. We first present the overall structure of AVASR systems with three main areas: audio feature extraction, visual feature extraction, and (subsequently) audio-visual speech integration. We present the MFCC features for an audio stream with the standard HMM modeling technique, then we describe the appearance- and shape-based visual features. Subsequently, we present two feature integration techniques, feature concatenation, and model fusion. We also discuss the results of a set of experiments conducted to select the best system setup for Polish under noisy audio conditions. The experiments simulate human-computer interaction in a computer control case with voice commands in difficult audio environments. With the Active Appearance Model (AAM) and multi-stream Hidden Markov Model (HMM), we can improve system accuracy by reducing the word error rate by more than 30% (as compared to audio-only speech recognition) when the signal-to-noise ratio drops to 0 dB.*

Keywords audio-visual speech recognition, visual feature extraction, human-computer interaction

Citation Computer Science 19(1) 2018: 41–63

1. Introduction

Speech processing is a key item in a natural human-computer interaction where this type of communication is highly integrated with a usage model, like virtual assistants or smart environments. This is even more important for all devices that are not equipped with 'standard' interaction mechanisms (e.g., mice, keyboards) due to their size (e.g., smart watches) or inability to use (e.g., car navigation systems). Automatic speech recognition (ASR) systems based on hidden Markov models (HMM) have become an industry-standard. Such systems perform very well under good audio conditions, but their performance decreases rapidly when the signal-to-noise (SNR) ratio drops. The other drawback is that HMM-based systems require training samples that match the testing/working conditions. When this requirement is not met, the system works unpredictably. One of the main challenges in the ASR domain is how to develop systems that work in real-world situations as well as they do in laboratory environments. This means robustness to all typical kinds of noise, like other sound sources (background, other people speaking, etc.), room reverberation, or microphone distortions. It may be especially difficult to secure good communication conditions in some public areas where spoken-language processing may be the most useful, like automatic information points (kiosks) on streets, airports, etc. One approach to this problem is to use a deep neural network-based model (DNN) that lead to some improvements in ASR accuracy; however, low SNR is still an issue. Another approach is to introduce another modality to complement the acoustic speech information that is not susceptible to acoustic distortion.

Human speech production and perception are bimodal in nature. This phenomenon was demonstrated by McGurk [28]: when the sound /ga/ is combined with video of a person uttering the sound /ba/, most people perceive the speaker as uttering the sound /da/. The most-important benefit from visual modality is the complimentary information about the place of articulation that can help disambiguate highly confusable acoustic units; for example, unvoiced consonants ('p' and 'k') or voiced consonants ('b' and 'd') [51]. The other benefits are supplying additional information to the audio speech segmentation and helping in audio source (speaker) localization.

Incorporating visual modality into an ASR system may help in overcoming difficult conditions and help generate a robust system [45]. By exploiting the visual modality of the speaker's mouth region, the automatic recognition of visual speech is formally known as *speechreading* or *lipreading*. Compared to audio-only speech recognition, AV-ASR introduces two challenging tasks. First, the visual-feature-extraction stage must be executed. This step requires robust face detection and region-of-interest (ROI) extraction and tracking. ROI mostly contains only the speaker's mouth region. ROI location estimation is followed by the visual-feature-extraction stage. Visual features may be divided into two main categories: appearance-based and shape-based. Visual features are also highly speaker-dependent and encode more information about a person's identity than his/her content of speech. To resolve this problem, some normalization techniques must be applied [23]. The second issue, an *audio-visual fusion*,

is the problem of looking for an efficient way to fuse the streams' information and avoid the situation when the system performance for combined streams is worse than one of its streams used independently. Audio and video information can be integrated by feature fusion, decision fusion, or model fusion. Feature fusion is a combination of audio and video features into a single vector, which is later processed by a single classifier. Features may simply be concatenated, reweighted (different domains), or processed with the PCA or LDA dimensionality reduction block to extract the most-important features. Decision fusion is based on independent stream processing in separate classifiers and a linear combination of class-belonging likelihoods. This fusion type provides a mechanism for modeling the reliabilities of each future stream. Stream reliability may vary during recognition; i.e., audio noise level may be increased, so the audio stream weight should be reduced. An efficient technique for data fusion is based on dynamic Bayesian networks [33]. These allow for data integration based on model fusion [50]. Additional improvements may be obtained when the DBN based model also incorporates constrained de-synchronization between modalities [48], while streams from the same modality should stay synchronized. The general scheme of an audio-visual speech recognition system is presented in Figure 1.

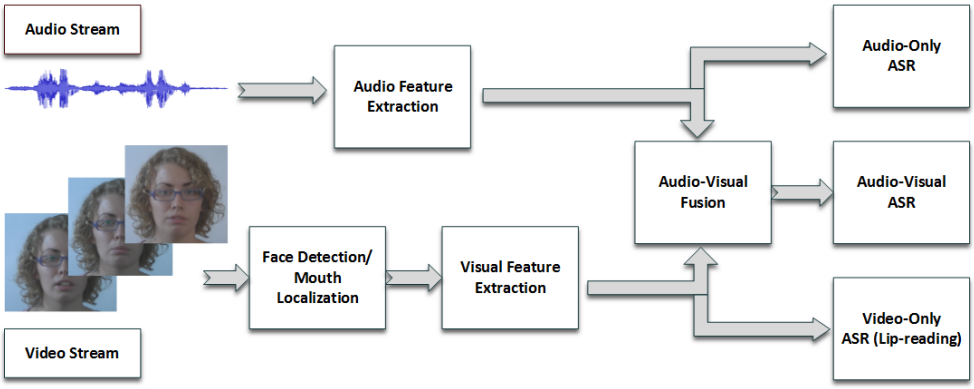

Figure 1. General scheme of audio-visual speech recognition system with separate tracks (without AV fusion) to audio-only and video-only (lip reading) speech recognition

In this work, we provide a brief overview of the main techniques applicable to Audio-Visual Automatic Speech Recognition and their potential in human-computer interaction for Polish speech. The results from some experiments are also presented. This paper is organized as follows: in Section 2, we describe previous work in the AVASR domain, then we present audio (Sec. 3) and visual stream (Sec. 4) processing methods, with ROI extraction and various parametrizations. Section 5 describes feature fusion techniques. The audio-visual database of Polish speech that was used for training as well as an evaluation of our system is presented in Section 6. The general setup of our AVASR system and the results from all of the experiments are described in Section 7. Final conclusions are presented in 8.

2. Related work

The first automatic speechreading system was reported in 1984 by Petajan [42]. Visual features were extracted from mouth ROI, interpolated by image thresholding. The following shape features were used: mouth height, width, perimeter, and area. The extracted features were used by a visual-only recognizer to rescore the best two hypothesis from an audio-only system. Since the Petajan work, many AVASR system have been presented that differ in four main aspects: the audio-visual database used during the training and system-evaluation phase, visual front-end design, audio-visual integration strategy, and speech recognition method used.

AVASR systems were tested on small vocabulary tasks with nonsense words [1], isolated words [5], connected letters [43], and digits [10]. Some attempts of continuous speech processing with small vocabularies have also been reported [25]. With new Audio-Visual corpora like AV-TIMIT [18] or XM2VTS [29], researchers have tried to run large-vocabulary continuous speech recognition with AVASR. Some experiments were based on stereo-vision and databases containing both frontal and profile speaker images [16].

Visual front-end differs in several points. Because the mouth area contains the most relevant information about what was spoken, the most-popular region-of-interest (ROI) is a rectangular area that contains only the speaker's mouth. In some cases, larger parts of the lower face (jaw, cheeks) [44] or even the entire face [35] were used. The next step after ROI extraction is the parametrization of the selected area. We can define three main feature types: shape-based, appearance-based, and a combination of these two. All geometric-type features that describe the mouth (the assumption that most of the information is coded in the contours of the speaker's lips) such as height, width, and area are used [1, 5]. Lip countour descriptors [9, 15] and statistical models of shape [24] were also investigated. The second category, appearance-based features, are becoming more popular [26]. Appearance features are built from various transforms of the whole ROI, because all pixels within an ROI may contain useful information. Nowadays, AVASR systems are based mostly on appearance features due to the lower cost of building initial models – training samples with manually annotated contours or positioned landmarks are not required. The final category is a concatenation of the shape and appearance features. Very popular is a statistical model named the active appearance model [6], which has been used in many systems.

The second aspect of a different AVASR system is the integration of information from the audio and visual streams. Audio-visual fusion is an instance of the general classifier combination problem where observations from two streams are available. The traditional approach to information-fusion schemes classifies them on early-, intermediate-, and late-fusion strategies. Another nomenclature is for feature-, classifier-, and decision-level fusion strategies, respectively.

The first one, feature-level fusion (early fusion), is also referred as the 'data to decision' fusion scheme [49]. In this case, one concatenates the feature vectors from the multiple modalities to obtain a combined feature vector. We can mention

two main advantages of this scheme: it provides better discriminatory abilities by exploiting covariations between the audio and video features, and it is much simpler to implement. The dimensionality of the resulting feature vector is often too large, so certain dimensionality reduction technique like discrete cosine transform (DCT), principal component analysis (PCA), or linear discriminant analysis (LDA) must be used. The main drawback of this fusion technique is that it cannot be applied to model asynchrony between different streams. It also performs poorly when the reliability of the different modalities during the training phase differ from the actual working phase. This fusion scheme was explored, for example, by [39, 27].

The coupled HMMs and the multistream HMMs exploited by [10] and dynamic Bayesian networks (DBNs, [14, 48]) are used in classifier-level fusion strategies. In such cases, the information is processed in a single classifier but with separate feature vectors. A composite classifier allows for the weighted combination of different modalities taken on each frame based on their reliability. Asynchrony between different streams can also be modeled to some extent. Both properties are very important for the audio--visual speech-processing task, where the audio and video asynchrony is of the order of 100 ms, whereas the frame duration is typically 25 ms [45]. In real-world situations, the reliability of the different streams varies with time. For example, the video channel might be completely unreliable if the speaker turns away from the camera, and audio stream reliability goes down when the background noise level increases.

The late (decision-level) fusion strategy involves the combination of probability scores or likelihood values obtained from separate unimodal classifiers for each stream based on some reliability weighting scheme. In this scheme, each classifier may be retrained with some additional data from a single domain. However, in the case of audio-visual speech recognition, it has been shown that this strategy gives worse results than model fusion [10].

One of the most-important trends in machine learning, deep neural networks (DNNs) have shown impressive performance in both audio and visual classification tasks. This mechanism was also introduced to AVASR systems at different levels: for feature extraction from different streams [38], where the multistream HMM was used as a classifier. DNNs were also used as complete classifiers with multimodal observations [37]. Using DNNs was also reported as beneficial not only in difficult audio conditions but even when the signal-to-noise ratio is high [32]. The main drawback with DNNs is that they require much more data for training, so it is especially difficult to build a DNN for under-resourced languages.

Most of the work in the audio-visual speech-processing domain has been made for English [14, 45], but some attempts in other languages have also been reported; e.g., Polish [22], French [10], Czech [40], or Japanese [34].

3. Audio stream processing

Recognizing speech directly from a digitized waveform is not possible due to the large variability between the same word utterances spoken even by a single person.

Automatic speech recognition systems that are based only on audio streams use one standard-feature-extraction scheme. The most-popular are Linear Predictive Coding (LPC)and Mel-frequency Cepstral Coefficients (MFCC). Some normalization and filtering techniques are also used for make the system more robust to the speaker and noise-level changes. In our audio-visual experiments, we are using an audio-processing front-end from the Sarmata ASR system [56] with MFCC parametrization. Input speech signal (16 kHz sampling rate, 16 bits/sample) is windowed first, with a Hamming Window of a length of 20 ms that is moved by a 10 ms offset. For each frame, an FFT transform is executed, resulting in 256 frequency bins containing spectrum magnitudes. Then, it is filtered with a set of 15 triangular filters that are equally spaced along the mel-scale defined in HTK book [55] (eq. 1) with lower- and upper-frequency cut-offs of 100 and 3900 Hz, respectively. The cepstral coefficients are calculated from log filterbank amplitudes m_j using the Discrete Cosine Transform (eq. 2), where N is the number of filterbank items. The final feature vector is reduced to $K = 12$, first coefficients c_k, and an energy feature. The energy is computed as a log of the signal energy; that is, for speech samples $s_n, n = 1, N$ (eq. 3). The energy coefficients may be normalized to overcome problems with underflow by applying a floor energy level and normalizing it to a range of $(-E_{min}; 1.0)$. To capture speech dynamics, each frame is concatenated with its first- and second-order derivatives, resulting in a 39-dimensional feature vector. This process is depicted in Figure 2. The other way to capture speech dynamics is an LDA projection of $\pm n$ consecutive frames of MFCC coefficients around the current frame. In particular, the effect of inserting a transmission channel on the input speech is to multiply the speech spectrum by the channel transfer function. In the log cepstral domain, this multiplication becomes a simple addition that can be removed by subtracting the cepstral mean from all input vectors. This technique (*Cepstral Mean Normalization*) is very effective in compensating for long-term spectral effects, like different microphones and between-speaker variability.

$$\mathrm{Mel}(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{1}$$

$$c(i) = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{2}$$

$$E = \log \sum_{n=1}^{N} s_n^2 \tag{3}$$

MFCCs are the parameterization of choice for many speech recognition applications. They give good discrimination and lend themselves to a number of manipulations. An alternative to Mel-Frequency Cepstral Coefficients is the use of Perceptual Linear Prediction (PLP) coefficients. The mel filterbank coefficients are weighted by an equal-loudness curve and then compressed by taking the cubic root. LP coefficients are estimated from the resulting auditory spectrum, which are then converted to cepstral coefficients in the normal way.

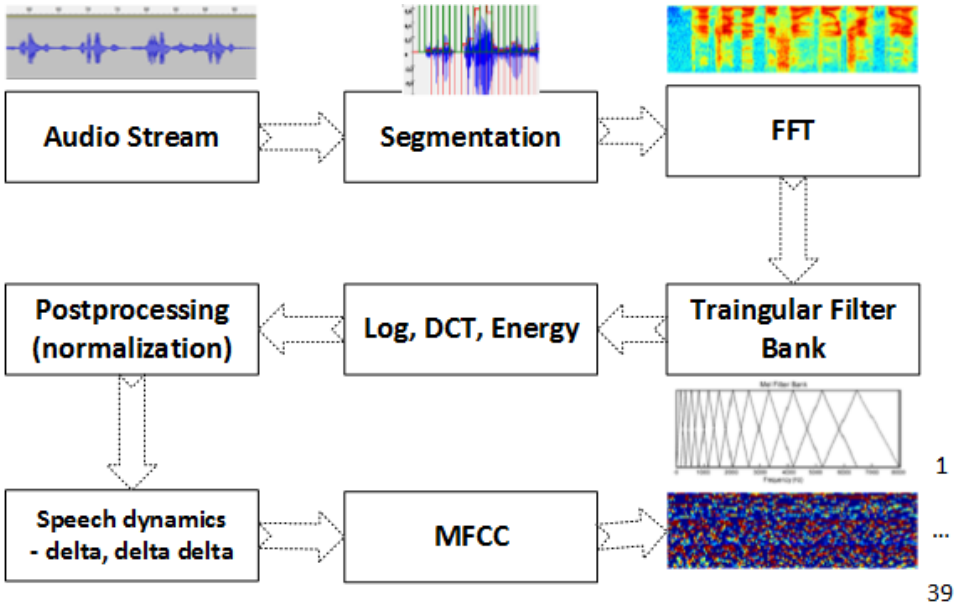

Figure 2. Data flow for audio stream processing during MFCC feature extraction

Basic units in speech recognition are phonemes. A phoneme is a minimal unit of sound that has a semantic content. Phonemes have some specific features by which their can be distinguished. Two major categories are vowels and consonants (which can be further separated to voiced and unvoiced). The consonants may be also separated by a manner of articulation to several classes, like stops, nasals, fricatives, and affricates. In the ASR system for Polish, we are using 37 phonetic classes and a silence model. The connection between words (ortographic notation) and their pronunciation (phonetic transcription) are defined in *dictionary*. The dictionaries may contain multiple transcriptions for each word, because there are several ways of how a word may be pronounced. The dictionaries are created by linguisticians or by specialistic software based on transcription rules [57].

One of the most-popular ways of modeling phones and connecting between them inside words is the *hidden Markov model* (HMM, see Fig. 3), where audio features extracted from an input stream are *observations* and phone labels are latent variables. Given a set of observation vectors $O = o_1, o_2, \ldots, o_T$, we are looking for word w_i that maximizes the score over equation 4.

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)} \tag{4}$$

For a given set of prior probabilities $P(w_i)$, the most-probable spoken word depends only on likelihood $P(O|w_i)$, which requires approximating joint conditional

probability $P(o_1, o_2, \ldots, o_T|w_i)$ by an estimation of the Markov model parameters for the most-likely state sequence (eq. 5).

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)}, \tag{5}$$

where $a_{x(t)x(t+1)}$ is the transition probability between states and $b_{x(t)}$ is the observation probability. When the state distribution probability is modeled by Gaussian Mixtures, the formula for computing $b_j(o_t)$ is then

$$b_j(o_t) = \sum_{m=1}^{M} c_{j_m} \frac{1}{\sqrt{2\pi|\Sigma_{jm}|}}, \exp\,(x - \mu_{jm})^T \Sigma_{jm}^{-1}(o_t - \mu_{jm}) \tag{6}$$

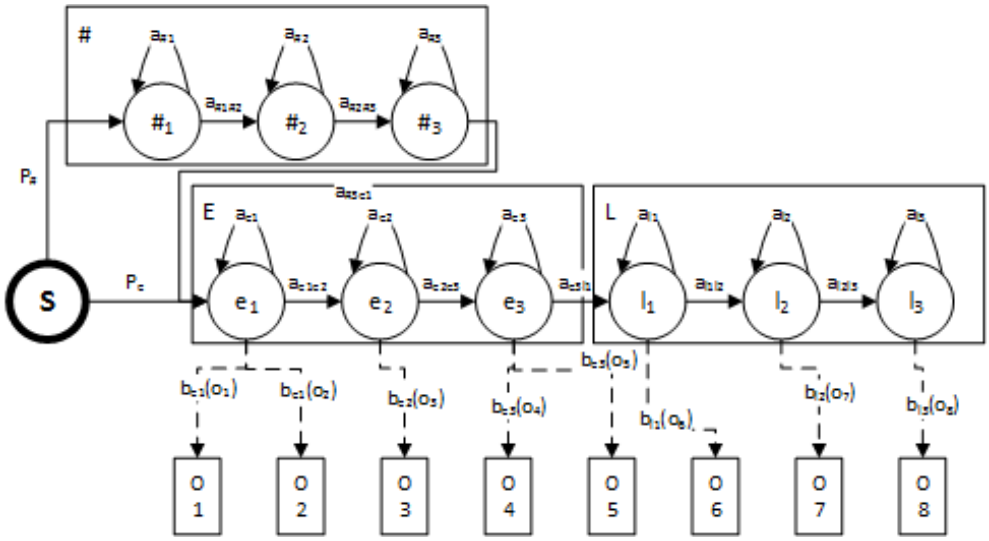

Figure 3. Scheme of building HMM word model by concatenating phone models, with optional silence at beginning

4. Video stream processing

The problem of visual-stream processing is two-fold: first, the speaker's face must be found on an image, and then the selected area is restricted to the region-of-interest (ROI) (containing only the mouth in most cases). The second step is a transformation of the extracted ROI to a relatively small number of informative features. Visual features may be broadly categorized in two approaches [23]. The first one (the top-down approach) is derived from higher-level, model-based features that utilize shape-only images or shape-and-appearance-of-mouth-area images and requires annotated samples during the learning stage. The second is a bottom-up approach, where the

model is built from low-level image-based features that are then compressed using one dimensionality-reduction algorithm, such as a discrete cosine transform (DCT) or principal component analysis (PCA).

The resulting visual features are often post-processed to improve system robustness and manage speaker variability. One of the most-important problems is that the feature extraction rate differs between the audio and visual streams. The large number of algorithms require that the features should be extracted uniformly for both modalities. This is often resolved by the simple element-wise linear interpolation of the visual features, which are extracted at 25 Hz to the audio frame rate, extracted at 100 Hz. The other problems are mainly related to the video stream. Between-speaker variability and recording conditions can be overcome by a subtraction of the vector mean over each utterance [55], which is a standard visual feature mean normalization (FMN) technique. In addition, utilizing visual speech dynamics [47] by augmenting static visual features with their first- and second-order temporal derivatives [55] may also be beneficial in improving recognition.

4.1. Region-Of-Interest extraction

Both types of features require the mouth-region extraction to be executed as an initial step. We are using the Viola-Jones algorithm [54], which is based on the idea of a boosted cascade of weak classifiers where each one is trained to accept a very large number of regions (high detection ratio and false positive but very low value of true-reject ratio). The square window moves over the whole image at different scales. All classifiers must accept the analyzed part of the image to regard this part as an interesting region (see Fig. 4).

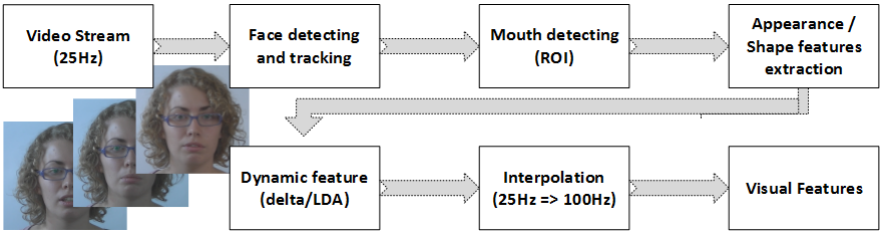

Figure 4. Data flow during visual speech feature extraction

As a first step, the speaker's face must be found in the image. After that, another run of the Viola-Jones algorithm is executed at the bottom-half of the face region with a mouth-trained classifier. The models defining the face and mouth area come from [4] and are provided with OpenCV framework. As a result, we gather a rectangular area with varying size, which is later processed. We are testing features from both categories. Visual features are extracted at 25 Hz and then interpolated to meet audio-feature frequency (100 Hz).

4.2. Appearance-based features

All pixels in the image part (mouth region typically, but sometimes larger portions of the lower face) are considered as informative for visual speech. The pixel values from the image converted to Hue-Saturation-Value color space or gray-scale are concatenated into a single vector. In the case of an $M \times N$-pixel region, the dimensionality of the resulting vector is too large. Some of the well-known image transformations must be used to obtain a feature vector that contains most speech-reading information within its $d \ll M \times N$ elements. The most-popular are principal component analysis (PCA), discrete cosine or wavelet transform (DCT, DWT), and Linear discriminant analysis (LDA). In our case, the appearance-based features (denoted as **DCT**) are extracted by downsampling the ROI to 64×32 pixels and converting it to a gray-scale image. A two-dimensional discrete cosine transform (DCT) is applied. The first 30 coefficients (without a DC value and their derivatives) constitute a 60-dimensional vector. The exemplary results from the following steps of appearance feature extraction are presented in Figure 5.

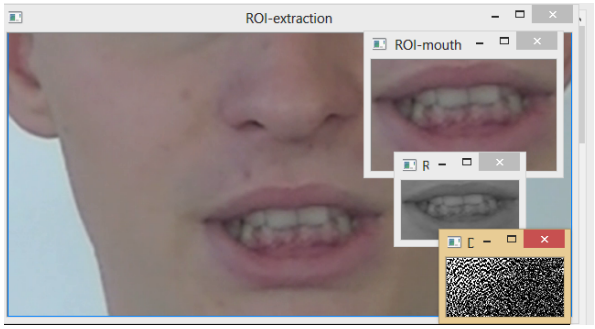

Figure 5. Mouth region and low-level image-based feature extraction (DCT)

4.3. Shape-based features

In contrast to the appearance-based features, the shape-based feature assumes that the contours (shapes) of the speaker's lips contain most of the speechreading information. The first attempts were based on geometric features, like lip width and height, area, and perimeter. Active shape model (**ASM**) is a statistical model that represents an object by the coordinates of a set of labeled points [7]. A number of K contour points (from the union of various face shapes; i.e., inner and outer lip) are first labeled on available training images, and their coordinates are placed on $2K$-dimensional vectors: $\mathbf{x} = [x_1, y_1, x_2, y_2, \ldots, x_K, y_K]$. A statistical model of the lip shape is identified by optimal orthogonal linear transform (PCA), given a set of vectors $\mathbf{x}$ that was tracked and aligned. Procrustes analysis is an iterative procedure used in mean shape $\bar{\mathbf{s}}$ computation. In our experiments, we are using a set of 24 points on the outer lips and 18 points along the inner lip contour. The first 12 components after PCA are used

as a feature vector. To obtain the shape features, an appropriate tracking algorithm is required. We are using the inverse compositional fitting algorithm [41] that was designed to fit the AAM features described below. Model adaptation is a two-stage process: first, the whole face model is fit to the image, and then the mouth-only model fitting starts from a face-initialized pose.

4.4. Combined features

The model-based features are using the Active Appearance Model (**AAM**) [6] algorithm that utilizes shape and appearance information. Shape component (7) is formed in the same way as in the ASM model, by concatenating the 2D coordinates of a set of n vertices that are boundary markers. Appearance vector (8) is defined by pixels that lie inside mean shape $\bar{\mathbf{s}}$. Appearance can be represented in a similar way as the shape component, as a base appearance and linear combination of k appearances (computed by applying PCA to the shape-normalized training images). The shape and appearance components are concatenated, reweighted (due to their natures: shape are coordinates and appearance are pixel intensities) and processed by a final PCA to obtain more-compact and decorrellated features.

$$s = \bar{s} + \sum_{i=1}^{m} p_i s_i \tag{7}$$

$$A = \bar{A} + \sum_{i=1}^{l} \lambda_i A_i \tag{8}$$

5. Fusion strategies

When more than one sensor is used to gather information, the features extracted from different modalities must be integrated in order to meet the sampling frequencies between streams (audio frames are sampled at 100 Hz, while video frames are at 25 Hz). Fusion schemes may be categorized in three main strategies: feature level, classifier level, and decision level. The first (also known as early fusion) is based on concatenation of the feature vectors from multiple modalities that are temporally correlated. A combined vector is then used for the classification task. This strategy can provide a better discriminatory ability for the classifier by exploiting the co-variations between the audio and video features [52]. The high-dimensionality of the feature vector requires the application of a dimensionality-reduction technique. Sometimes, it is beneficial to run hierarchical dimensionality reduction [45].

Feature concatenation is the simplest method to implement (see Fig. 6). Concatenated features may be processed by systems that have previously worked for audio-only speech recognition. However, it cannot model asynchrony between the streams well, nor can it handle different reliabilities of the modalities during the training and testing phases.

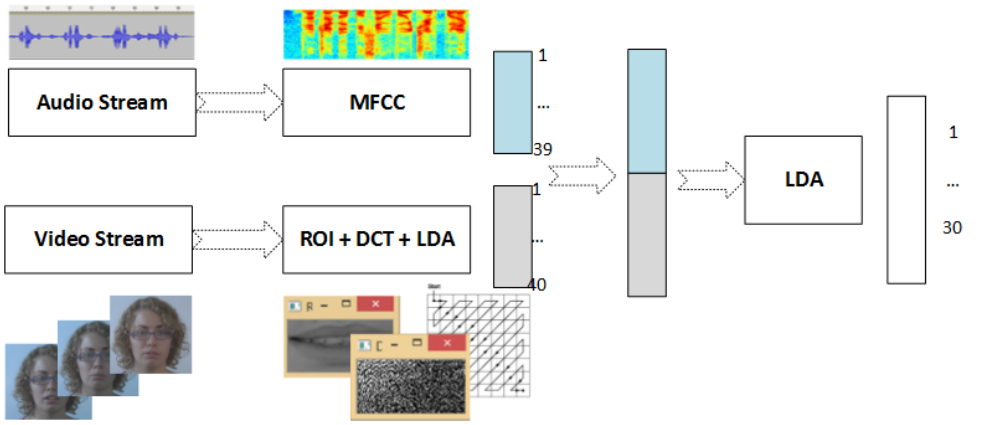

Figure 6. Feature concatenation

In the classifier-fusion strategy, the information is fused within a single classifier, but only after processing the features separately (a composite classifier, like the multistream Hidden Markov Model or Dynamic Bayesian Network may be used, see Fig. 7). At this level, a weighted combination of different modalities may be used; some asynchrony between streams may also be modeled. The late (or decision) fusion strategy involves a combination of the likelihood scores obtained from separate classifiers that run on each stream to get a combined decision. In this approach, the reliability of the streams is introduced by an exponentially weighting linear combination of likelihoods from independent classifiers. In AVASR systems, this strategy has been shown to give worse results than both early and intermediate fusion [10].

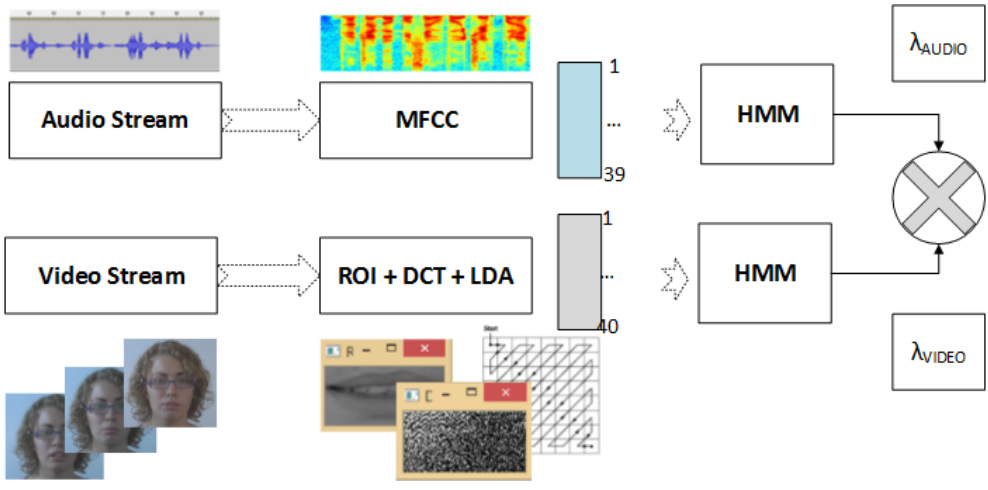

Figure 7. Decision fusion strategy

In our experiments, we are using two fusion schemes: feature concatenation followed by LDA projection for dimensionality reduction, and model-level fusion with two-stream, left-to-right, three-state HMM models defined for each phoneme (due to the nature of multi-stream HMM classifier, separate viseme models for the visual stream are not extracted; we are using 37 phonemes and a silence model for both streams) and concatenated to model whole words.

Multistream HMM

Multistream HMM (MSHMM) is an HMM-based structure that handles multiple modalities for timeline data where we can model synchronous and independent streams (see Fig. 8). For the continuous case, multi-stream HMM was originally introduced to fuse the audio and visual streams in speech recognition using continuous HMM [20]. In this case, independent streams are treated separately, the feature space is partitioned into subspaces, and different probability density functions (pdf) are learned for the corresponding streams. The relevance of the different streams is encoded by exponent weights, and a weighted geometric mean of the streams is used to approximate the pdf.

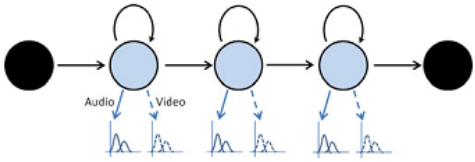

Figure 8. Example of a multistream HMM with audio and video streams and three states per stream

To learn all of the model parameters, a two-step learning mechanism is employed. In the first step, the standard Baum-Welch algorithm [46] is used to learn all model parameters. In the second step, discriminative training is used to learn the exponent weights of adequate streams. The main drawback of this approach is its inability to provide an optimization framework that learns all of the HMM parameters simultaneously. In addition, solving this issue using two layers of training that optimize two different types of parameters is susceptible to local optima. To alleviate these limitations, the authors in [31] proposed an MSHMM structure that allows for the simultaneous learning of all model parameters (including the stream-relevance weights) by linearizing the approximation of the pdf. In this approach, the stream-relevance weights were introduced at the mixture level, and the Baum-Welch (BW) learning algorithm was generalized to derive the necessary conditions to learn all parameters simultaneously. Compared to HMM, the observation probability $b_{x(t)}$ is modified to

$$b_j(o_t) = \prod_{k=1}^{L} \left[\sum_{m=1}^{M} c_{j_m k} \frac{1}{\sqrt{2\pi|\Sigma_{jmk}|}}, \exp\left(x - \mu_{jmk}\right)^T \Sigma_{jmk}^{-1} (o_t - \mu_{jmk}) \right]^{w_{jk}} \tag{9}$$

where $\sum_{k=1}^{L} w_{jk} = 1$.

This is known as state-level weighting. Total observation probability is weighted product w_{jk} of the observation probabilities from K separate streams. The weights are learned during training and are adjusted to each phone/state separately according to its reliability and relevance.

Dynamic Bayesian network

The Bayesian network encodes the dependencies between sets of random variables, which are represented as edges and nodes of a directed graph. A dynamic Bayesian network is an extension of a plain network that is used for modeling random variable evolution over time. It is achieved by repeating the network structure and connecting the corresponding nodes. In speech recognition, nodes in a network represent hidden variables (words, phonemes, transitions) and observed variables (acoustic features like MFCC), while the edges correspond to conditional probability functions or deterministic dependencies [3]. For example, when modeling word-inner phone position and phone transition, the next phone is determined by word transcription and the occurrence of phone transition in a previous time-slice. Random transition may be used for incorporating language models.

The structure of the model used in our system is shown in Figure 9. Observable features are presented as filled circles. For audio streams, the features extracted with MFCC are marked as $\mathrm{O}^{\mathrm{MFCC}}$, $\mathrm{O}^{\mathrm{DWT}}$ and are wavelet features. Features from the speaker's mouth region used in video stream are labeled with O^{V}. For modeling asynchrony between modalities, there are two different nodes that represents the *phoneme* and its *states*, but both modalities share a single *word* variable. The audio signal is parametrized with two independent algorithms, but both audio streams are synchronized on the same *state* of the phoneme.

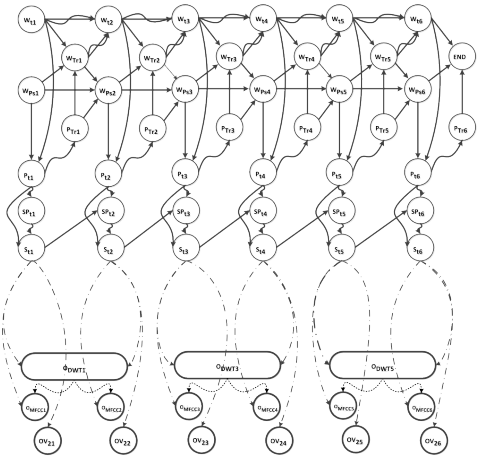

Figure 9. Scheme of DBN models with two modalities and asynchrony represented by separate phone nodes

6. Audio-visual corpus of Polish speech

In this work, we have used the audio-visual database of Polish speech recordings [21]. It contains three different types of utterances: numbers and commands (160 short phrases from a list of the most-popular questions asked to Virtual Assistant), using spoken language, and parts of read texts (7 different utterances, like articles, definitions, and part of stories). Longer utterances was split into several-words phrases for learning and testing purposes. There are 24 speakers: 13 males and 11 females. The recordings contain only faces (frontal view) on bright background with rather-invariant lighting conditions, and feature full HD quality with 25 frames per second. Each speaker has been recorded for about 10 minutes, totaling about 4 hours.

7. Experiments

In our testing environment, the audio-visual database was divided into two sets. The training set consisted of 21 speakers, and the 3 remaining speakers were used as the testing set. This was repeated 8 times to test all speakers, resulting in about 1000 testing utterances from the human-computer interaction domain. All remaining data was used for system training. In all of the following experiments, an original audio stream was mixed with random samples of background noise from various environments at eight different SNR levels (from 30 dB to 0 dB). The system performance for clean audio conditions was also investigated. The background noise was a random part of the CHiME-3 noise database [2]. We are using data recorded in one channel (from the six available) and for all four environments where the background noise recordings were collected: in a cafe, at a street junction, in public transport, and at a pedestrian area.

7.1. System implementation

The Audio-Visual Automatic Speech Recognition (AVASR) system is based on the ASR System for Polish, called *Sarmata* [56]. *Sarmata* uses MFCC for audio parametrization. It works on HMM models and context-dependent phoneme representations. The audio-visual extension is implemented in C# .NET language, with the use of the *EmguCV* framework for image processing and computer vision algorithms. *EmguCV* is a *.NET* platform wrapper for the well-known C++ library *OpenCV*. When the feature concatenation method was used for data fusion, AVASR extensions were used during the feature extraction stage for enhancing the MFCC features according to previous descriptions. Further processing of the feature vector was done by a standard ASR engine and HMM decoder. For the model fusion technique, both stages (feature extraction and decoding) were performed inside the AVASR code. For multi-stream HMM, our own implementation was used, while for DBN modeling, we used an *Infer.NET* framework.

7.2. Results

The first task was to compare different visual stream parametrization methods under various noisy conditions when feature concatenation was used as a fusion scheme. The collected results are presented in Figure 10.

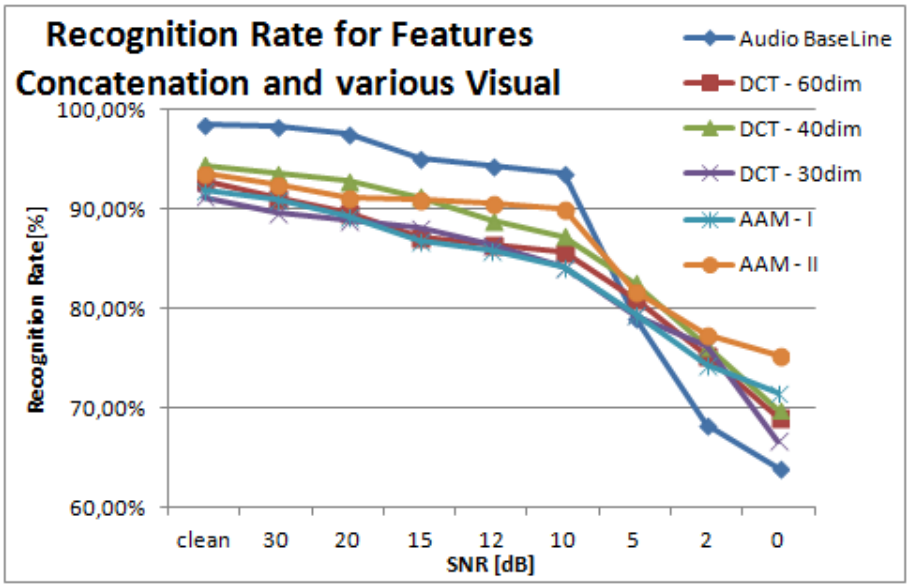

Figure 10. Comparison of different visual feature extraction methods

Three different AVASR setups with appearance-based visual features (DCT) and two with Active Appearance Models (AAM) that differ between the corresponding setups by visual feature vector dimensionality. For the DCT feature after LDA transformation, the resulting vector was investigated for 30, 40, and 60 dimensions. The first AAM model setup was built from 42 points along the contours of the mouth (24 points along the outer lip and 18 points from the inner lip). The top 12 eigenshapes were used as a shape component of AAM. The shape features were combined with the appearance component, resulting in a 40-dimensional vector after the LDA reduction of 5 concatenated frames (to gather speech dynamics). All AVASR features were compared with the results from an audio-only speech recognition system. We can see that, when audio conditions are very good, it is hard for the AVASR system with feature concatenation as a fusion scheme to approach the audio-only speech recognition system, regardless of the visual feature setup. Two aspects might be important for the most part: the audio system was trained on a large amount of mostly clean data (near 100 h), and when the testing conditions match the training conditions, system performance is very high. In this situation, the visual stream gives no additional information to the audio stream, and it is harder to train a system with less

data when the feature vector has more components. When SNR goes down, the visual components becomes more and more beneficial. The appearance-only feature (DCT) gives worse results than both the AAM models, and the second AAM setup is the most stable regarding noisy audio conditions, giving an over 75% recognition rate at 0 dB (24.76% Word Error Rate, which is more than a 30% reduction from 36% WER for audio-only ASR). The best results for DCT was achieved for the 40-dim vector (30.16% WER at 0 dB).

In the second experiment, the best visual features selected from the previous experiment (one for AAM and one for DCT) were used to test the feature fusion schemes. The AAM model with N parametrization points and 40-dim DCT vector were used to investigate the multistream Hidden Markov Model (HMM) and Dynamic Bayesian Network (DBN) classifiers apart from feature concatenation fusion (FC). Nearly all of the new setups give better results than plain feature concatenation at clean audio conditions (except the DBN with DCT visual features); however, neither of them outperforms the ASR system.

The transition between 10 and 5 dB SNR is still critical to recognize the incorporation of the visual stream to audio stream processing as beneficial. Multistream HMM gives better results than both DBN and feature concatenation. The active appearance model is still more stable than the DCT features.

The results are presented in Figure 11, Table 1 and Table 2.

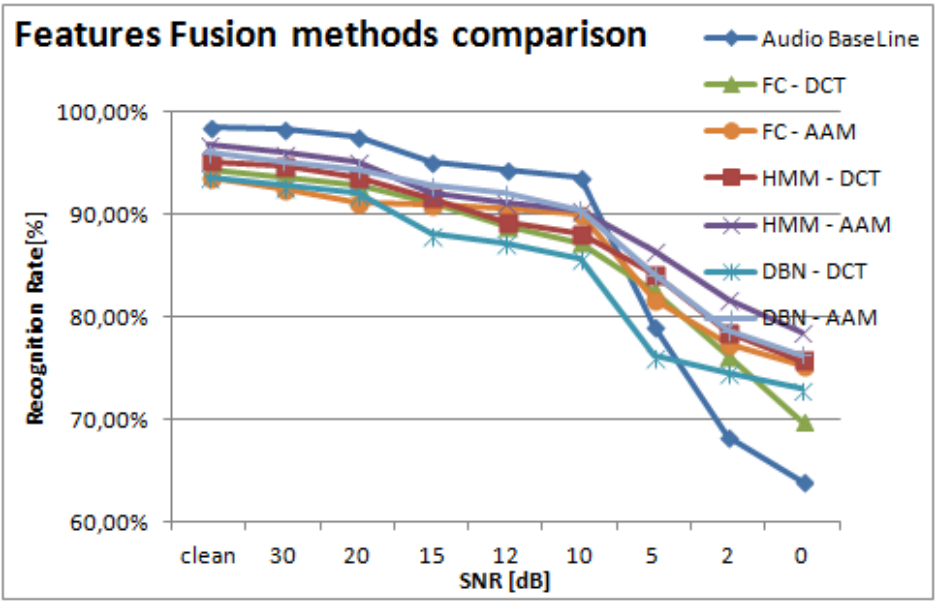

Figure 11. Comparison of different feature fusion schemes

Table 1
Phrase recognition rates for various visual features and feature concatenation fusion and selected SNR

Experiment	0 dB	2 dB	5 dB	10 dB	15 dB	20 dB	Clean
Audio [%]	63.99	68.32	79.17	93.65	95.24	97.62	98.51
DCT-60dim [%]	69.05	75.40	80.95	85.71	87.30	89.68	92.86
DCT-40dim [%]	69.84	76.19	82.54	87.30	91.27	92.86	94.44
DCT-30dim [%]	66.67	76.19	79.37	84.13	88.10	88.89	91.27
AAM-I [%]	71.59	74.44	79.44	84.29	86.90	89.37	91.90
AAM-II [%]	75.24	77.38	81.67	90.08	91.03	91.27	93.57

Table 2
Phrase recognition rates for different feature fusion methods with best visual features, at selected SNR

Experiment	0 dB	2 dB	5 dB	10 dB	15 dB	20 dB	Clean
Audio [%]	63.99	68.32	79.17	93.65	95.24	97.62	98.51
FC-DCT [%]	69.84	76.19	82.54	87.30	91.27	92.86	94.44
FC-AAM [%]	75.24	77.38	81.67	90.08	91.03	91.27	93.57
HMM-DCT [%]	75.79	78.57	84.13	88.10	91.67	93.56	95.24
HMM-AAM [%]	78.57	81.75	86.51	90.48	92.06	95.24	96.83
DBN-DCT [%]	73.02	74.60	76.19	85.71	88.10	92.06	93.65
DBN-AAM [%]	76.53	78.65	84.13	90.48	92.86	94.44	96.03

8. Conclusions

In this paper, we provided a brief literature review of the basic components in the automatic processing of audio-visual speech signals. The main three components of an AVASR system were described; namely, the audio signal extraction and processing, visual stream processing with extraction of an interesting region from a whole-face image, and different types of ROI parametrization. The strategies for audio-visual feature integration with multi-stream hidden Markov models and Dynamic Bayesian Networks were also presented. We focused on the algorithms and techniques used in our audio-visual speech-processing system.

In the second part, we also presented the results of audio-visual speech recognition under noisy conditions as compared to the results of an audio-only speech recognition system. The experiments were evaluated using an audio-visual database of Polish speech, where additional noise was introduced to the audio channel with different intensities. We selected simple commands and single words from the database to simulate human-computer interaction for the computer control case. The experiments show that it may be beneficial to add visual modality to speech processing, especially in low signal-to-noise conditions, to reduce word error rate by more than 30%. It is important to note that, for processing visual modality, we need much more computational power than for audio-only speech processing. The experiments

show that using an audio-visual speech-processing system for recognizing the Polish language may be reasonable for human-computer interaction, especially under some difficult audio conditions; i.e., in public places where SNRs are low (which was also described in this work).

Acknowledgements

This work was supported by Polish National Science Centre (NCN) granted by decision DEC-2011/03/N/ST7/00443.

References

[1] Adjoudani A., Benoit C.: On the integration of auditory and visual parameters in an HMM-based ASR. In: *Speechreading by Humans and Machines*, pp. 461–471, Springer, 1996.

[2] Barker J., Marxer R., Vincent E., Watanabe S.: The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In: *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 504–511, 2015. http://dx.doi.org/10.1109/ASRU.2015.7404837.

[3] Bilmes J., Bartels C.: Graphical model architectures for speech recognition, *IEEE Signal Processing Magazine*, vol. 22(5), pp. 89–100. http://dx.doi.org/10.1109/MSP.2005.1511827.

[4] Castrillón M., Déniz O., Guerra C., Hernández M.: ENCARA2: Real-time detection of multiple faces at different resolutions in video streams, *Journal of Visual Communication and Image Representation*, pp. 130–140, 2007.

[5] Chan M.T., Zhang Y., Huang T.S.: Real-time lip tracking and bimodal continuous speech recognition. In: *Multimedia Signal Processing, 1998 IEEE Second Workshop on*, pp. 65–70, 1998.

[6] Cootes T.F., Edwards G.J., Taylor C.J.: Active appearance models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(6), pp. 681–685, 2001. http://dx.doi.org/10.1109/34.927467.

[7] Cootes T.F., Taylor C.J., Cooper D.H., Graham J.: Active shape models-their training and application, *Computer vision and image understanding*, vol. 61(1), pp. 38–59, 1995.

[8] Cox S.J., Harvey R., Lan Y., Newman J.L., Theobald B.J.: The challenge of multispeaker lip-reading. In: *AVSP*, pp. 179–184, Citeseer, 2008.

[9] Czap L.: Lip representation by image ellipse. In: *The Proceedings of the 6th International Conference on Spoken Language Processing (Volume IV)*, 2000.

[10] Dupont S., Luettin J.: Audio-visual speech modeling for continuous speech recognition, *IEEE Transactions on Multimedia*, vol. 2(3), pp. 141–151, 2000. http://dx.doi.org/10.1109/6046.865479.

[11] Frankel J., Wester M., King S.: Articulatory feature recognition using dynamic Bayesian networks, *Computer Speech & Language*, vol. 21(4), pp. 620–640, 2007. http://dx.doi.org/10.1016/j.csl.2007.03.002.

[12] Gałka J., Ziółko M.: Wavelet parameterization for speech recognition. In: *Proceedings of an ISCA tutorial and research workshop on non-linear speech processing NOLISP 2009, VIC*, 2009.

[13] Gopinath R.A.: Maximum likelihood modeling with Gaussian distributions for classification. In: *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2, pp. 661–664, 1998.

[14] Gowdy J., Subramanya A., Bartels C., Bilmes J.: DBN based multi-stream models for audio-visual speech recognition. In: *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 1, pp. I-993–I-966, 2004. http://dx.doi.org/10.1109/ICASSP.2004.1326155.

[15] Gurbuz S., Tufekci Z., Patterson E., Gowdy J.N.: Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition. In: *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 1, pp. 177–180, 2001.

[16] Harte N., Gillen E.: TCD-TIMIT: An audio-visual corpus of continuous speech, *IEEE Transactions on Multimedia*, vol. 17(5), pp. 603–615, 2015.

[17] Hazen T.J.: Visual model structures and synchrony constraints for audio-visual speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14(3), pp. 1082–1089, 2006. http://dx.doi.org/10.1109/TSA.2005.857572.

[18] Hazen T.J., Saenko K., La C.H., Glass J.R.: A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In: *Proceedings of the 6th international conference on Multimodal interfaces*, pp. 235–242, 2004.

[19] Hermansky H.: Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, vol. 87(4), pp. 1738–1752, 1990.

[20] Hernando J.: Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition. In: *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97. 1997 IEEE International Conference on*, vol. 2, pp. 1267–1270, 1997.

[21] Igras M., Ziółko B., Jadczyk T.: Audiovisual database of Polish speech recordings, *Studia Informatica*, vol. 33(2B), pp. 163–172, 2013.

[22] Kubanek M., Bobulski J., Adrjanowicz L.: Lip tracking method for the system of audio-visual Polish speech recognition. In: *International Conference on Artificial Intelligence and Soft Computing*, pp. 535–542, Springer, 2012.

[23] Lan Y., Theobald B.J., Harvey R., Ong E.J., Bowden R.: Improving visual features for lip-reading. In: *Proceedings of International Conference on Auditory-Visual Speech Processing*, pp. 142–147 2010.

[24] Luettin J., Thacker N.A.: Speechreading using probabilistic models, *Computer Vision and Image Understanding*, vol. 65(2), pp. 163–178, 1997.

[25] Marcheret E., Libal V., Potamianos G.: Dynamic Stream Weight Modeling for Audio-Visual Speech Recognition. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–945–IV–948, 2007. http://dx.doi.org/10.1109/ICASSP.2007.367227.

[26] Matthews I., Potamianos G., Neti C., Luettin J.: A comparison of model and transform-based visual features for audio-visual LVCSR. In: *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, pp. 825–828, 2001. http://dx.doi.org/10.1109/ICME.2001.1237849.

[27] McCowan I., Gatica-Perez D., Bengio S., Lathoud G., Barnard M., Zhang D.: Automatic analysis of multimodal group actions in meetings, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(3), pp. 305–317, 2005.

[28] McGurk H., MacDonald J.: Hearing lips and seeing voices, *Nature*, vol. 264, pp. 746–748, 1976. http://dx.doi.org/10.1038/264746a0.

[29] Messer K., Kittler J., Sadeghi M., Marcel S., Marcel C., Bengio S., Cardinaux F., Sanderson C., Czyz J., Vandendorpe L., et al.: Face verification competition on the XM2VTS database. In: *Audio-and Video-Based Biometric Person Authentication*, pp. 964–974, Springer, 2003.

[30] Minka T., Winn J., Guiver J., Webster S., Zaykov Y., Yangel B., Spengler A., Bronskill J.: Infer.NET 2.6, 2014. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

[31] Missaoui O., Frigui H.: Optimal feature weighting for the continuous HMM. In: *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pp. 1–4, 2008.

[32] Mroueh Y., Marcheret E., Goel V.: Deep Multimodal Learning for Audio-Visual Speech Recognition. In: *arXiv preprint arXiv:1501.05396*, 2015.

[33] Murphy K.P.: *Dynamic bayesian networks: representation, inference and learning.* Ph.D. thesis, University of California, Berkeley, 2002.

[34] Nakamura S., Ito H., Shikano K.: Stream weight optimization of speech and lip image sequence for audio-visual speech recognition. In: *Sixth International Conference on Spoken Language Processing, ICSLP 2000*, pp. 20–24, 2000.

[35] Neti C., Potamianos G., Luettin J., Matthews I., Glotin H., Vergyri D., Sison J., Mashari A.: Audio visual speech recognition. Tech. rep., IDIAP, 2000.

[36] Newman J.L., Cox S.J.: Automatic visual-only language identification: A preliminary study. In: *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 4345–4348. IEEE, 2009.

[37] Ngiam J., Khosla A., Kim M., Nam J., Lee H., Ng A.Y.: Multimodal deep learning. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696, 2011.

[38] Noda K., Yamaguchi Y., Nakadai K., Okuno H.G., Ogata T.: Audio-visual speech recognition using deep learning, *Applied Intelligence*, vol. 42(4), pp. 722–737, 2015.

[39] O'Donovan A., Duraiswami R., Neumann J.: Microphone arrays as generalized cameras for integrated audio visual processing. In: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, 2007.

[40] Palecek K., Chaloupka J.: Audio-visual speech recognition in noisy audio environments. In: *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, pp. 484–487, 2013.

[41] Papandreou G., Maragos P.: Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, 2008.

[42] Petajan E.D.: *Automatic lipreading to enhance speech recognition (speech reading)*. Ph.D. thesis, University of Illinois at Urbana-Champaign, 1984.

[43] Potamianos G., Graf H.P.: Discriminative training of HMM stream exponents for audio-visual speech recognition. In: *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 6, pp. 3733–3736, 1998.

[44] Potamianos G., Neti C.: Improved ROI and within frame discriminant features for lipreading. In: *Image Processing, 2001. Proceedings. 2001 International Conference on*, vol. 3, pp. 250–253, 2001.

[45] Potamianos G., Neti C., Gravier G., Garg A., Senior A.: Recent advances in the automatic recognition of audiovisual speech, *Proceedings of the IEEE*, vol. 91(9), pp. 1306–1326, 2003. http://dx.doi.org/10.1109/JPROC.2003.817150.

[46] Rabiner L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. In: *Proceedings of the IEEE*, vol. 77(2), pp. 257–286, 1989.

[47] Rosenblum L.D., Saldaña H.M.: Time-varying information for visual speech perception. In: Campbell R., Dodd B., Burnham D. (eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*, pp. 61–81, Psychology Press, Erlbaum (UK), Taylor & Francis, 1998.

[48] Saenko K., Livescu K.: An asynchronous DBN for audio-visual speech recognition. In: *Spoken Language Technology Workshop, 2006. IEEE*, pp. 154–157, 2006. http://dx.doi.org/10.1109/SLT.2006.326841.

[49] Schwartz J.L., Robert-Ribes J., Escudier P.: Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. In: Campbell R., Dodd B., Burnham D. (eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech*, pp. 85–108, Psychology Press, Erlbaum (UK), Taylor & Francis 1998.

[50] Shivappa S.T., Trivedi M.M., Rao B.D.: Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey, *Proceedings of the IEEE*, vol. 98(10), pp. 1692–1715, 2010.

[51] Summerfield A.Q.: Some preliminaries to a comprehensive account of audio-visual speech perception. In: Dodd B., Campbell R. (eds.), *Hearing by eye: The psychology of lip-reading*, pp. 3–51, Lawrence Erlbaum Associates, Inc, 1987.

[52] Teissier P., Robert-Ribes J., Schwartz J.L., Guérin-Dugué A.: Comparing models for audiovisual fusion in a noisy-vowel recognition task, *IEEE Transactions on Speech and Audio Processing*, vol. 7(6), pp. 629–642, 1999.

[53] Tremain T.E.: The government standard linear predictive coding algorithm: LPC-10, *Speech Technology*, vol. 1(2), pp. 40–49, 1982.

[54] Viola P., Jones M.: Rapid object detection using a boosted cascade of simple features. In: *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. 1–9, 2001.

[55] Young S., Evermann G., Gales M., Hain T., Kershaw D., Liu X.A., Moore G., Odell J., Ollason D., Povey D., et al.: *The HTK book (for HTK version 3.4)*, Cambridge University Engineering Department, 2006.

[56] Ziółko M., Gałka J., Ziółko B., Jadczyk T., Skurzok D., Mąsior M.: Automatic speech recognition system dedicated for Polish. In: *Proceedings of the INTERSPEECH 2011 Conference, Florence, Italy*, pp. 3315–3316, 2011.

[57] Ziółko M., Ziółko B., Skurzok D.: Ortfon2 – tool for orthographic to phonetic transcription. In: *Human language technologies as a challenge for computer science and linguistics: Proceedings of 7th language & technology conference, November 27–29, 2015, Poznań, Poland*, pp. 115–119, 2015.

Affiliations

Tomasz Jadczyk
AGH University of Science and Technology, Faculty of Computer Science, Electronics and Telecommunications, Department of Computer Science, Krakow, Poland,
jadczyk@agh.edu.pl.pl

Received: 20.02.2017
Revised: 11.12.2017
Accepted: 11.12.2017