*Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015)*
*Barcelona, Spain, July 13 - 14, 2015*
*Paper No. 343*

# Audio-Visual Speech Processing System for Polish with Dynamic Bayesian Network Models

**Tomasz Jadczyk, Mariusz Ziółko**
AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
www.dsp.agh.edu.pl,
Techmo sp z o.o., techmo.pl
Kraków, Poland,
jadczyk@agh.edu.pl; ziolko@agh.edu.pl

***Abstract***- In this paper we describe a speech processing system for Polish which utilizes both acoustic and visual features and is based on Dynamic Bayesian Network (DBN) models. Visual modality extracts information from speaker lip movements and is based alternatively on raw pixels and discrete cosine transform (DCT) or Active Appearance Model (AAM) features. Acoustic modality is enhanced by using two parametrizations - standard MFCC and discrete wavelet transform (DWT) features, extracted for frequency subbands designed in accordance to human perception. An asynchrony between audio and visual modalities is also represented within the DBN model. Our system is evaluated on Audiovisual Speech Corpus of Polish, which contains most popular phrases from polish dialogs, uttered by 24 speakers. Experiments under various - clean and noisy - conditions confirmed that using double-parametrized audio and visual modalities (either DCT or AAM) with DBN model can reduce the Phrase Recognition Error Rate up to 35% under low signal-to-noise (SNR) conditions, comparing to results from standard audio-only, Hidden Markov Model - based system with MFCC parametrization.

***Keywords:*** audio-visual speech recognition, visual features extraction, Dynamic Bayesian Networks

## 1 Introduction

Speech processing is a key item in natural human-computer interaction. Automatic speech recognition (ASR) systems based on a hidden Markov models (HMM) became an industry-standard. Such systems perform well in good noise conditions, but their performance decreases rapidly when signal-to-noise (SNR) ratio falls down. Using deep neural networks (DNN) lead to some improvements in ASR accuracy, but low SNR is still an issue. It may be especially difficult to secure good communication conditions in some public areas, where spoken language processing may be especially applicable, like automatic information points (kiosks) on streets, airports, etc. Incorporating visual modality into ASR system may help in overcoming difficult conditions and generation of robust system, as shown by (Potamianos et al. 2003). Additionally, using multiple parametrizations for the same modality of speech may also lead to further improvements (Gowdy et al. 2004). The main goal when using multiple streams is finding efficient way for streams fusion, to avoid situation when system performance for combined streams is worse than one of its stream used independently. An efficient technique for data fusion is based on dynamic Bayesian networks (Murphy 2002). It allows data integration based on model fusion (Shivappa et al. 2010). Additional improvements may be obtained, when DBN based model also incorporates constrained de-synchronization between modalities (Saenko and Livescu 2006), while streams from the same modality should stay synchronized.

In this work, we propose the use of DBN algorithm and DBN based models for combining information from enhanced, double parametrized audio stream and additional visual stream. Moreover, our model is able to represent some level of asynchrony between audio and visual modalities. This paper is organized as follow: in Section 2 we describe general DBN algorithm and present network used in our system., section 3 presents scheme of our system, data flow and parametrization methods for each stream, then the database used for training and evaluation is described. Results are described in section 4, then final conclusions are presented in section 5.

## 2 Dynamic Bayesian Network

Bayesian network encodes dependencies between set of random variables, which are represented as edges and nodes of a directed graph respectively. Dynamic Bayesian network is an extension of plain network, used for modeling random variables evolution over time. It is achieved by repeating network structure and connecting corresponding nodes. In speech recognition, nodes in network represents hidden (words, phonemes, transitions) and observed (acoustic features, like MFCC) variables, while edges corresponds to conditional probability functions or deterministic dependencies (Bilmes and Bartels 2005). For example, when modeling word-inner phone position and phone transition, the next phone is determined by word transcription and occurrence of phone transition in previous time-slice. Random transition may be used for incorporating language models.

### 2.1 DBN models used in AVASR system

The structure of model used in our system is shown in Figure 1. Observable features are presented as filled circles. For audio streams, features extracted with MFCC are marked as $O^M$, $O^D$ 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 *phone* in audio, and *viseme* in visual modality, but both modalities share single *word* variable W. In the same time slot, W depends on both P and V variables, while word transition TR depends on a phone transition $P^{Tr}$ and viseme transition $V^{Tr}$ respectively. Audio signal is parametrized with two, independent, algorithms, but both audio streams are synchronized on the same *phone*.

## 3 AVASR SYSTEM

The AVASR system is based on our standard ASR System for Polish, called *Sarmata* (Ziółko et al. 2011). ASR system works on HMM models and context-dependent phoneme representations. The audio-visual extension is implemented in C# .NET language, with using *EmguCV* framework (Web-1) for image processing and computer vision algorithms, and *Infer.NET* framework (Minka et al. 2014) for DBN modeling. *EmguCV* is a *.NET* platform wrapper for a well known C++ library *OpenCV*.

### 3.1 Data flow scheme

The general data flow in our system is presented in Figure 2. First part, Voice Activity Detection (VAD) is based only on audio stream processing and is used for computational reduction. Audio stream is parametrized with two separate algorithms. For video stream, region-of-interest that contains mouth only area is extracted first, then it is parametrized according to Active Appearance Models algorithm or simple pixel intensities from ROI, with DCT dimensionality reduction (described in sec. 3.3). All features are passed to DBN algorithm, where inferring about word hypothesis is executed.

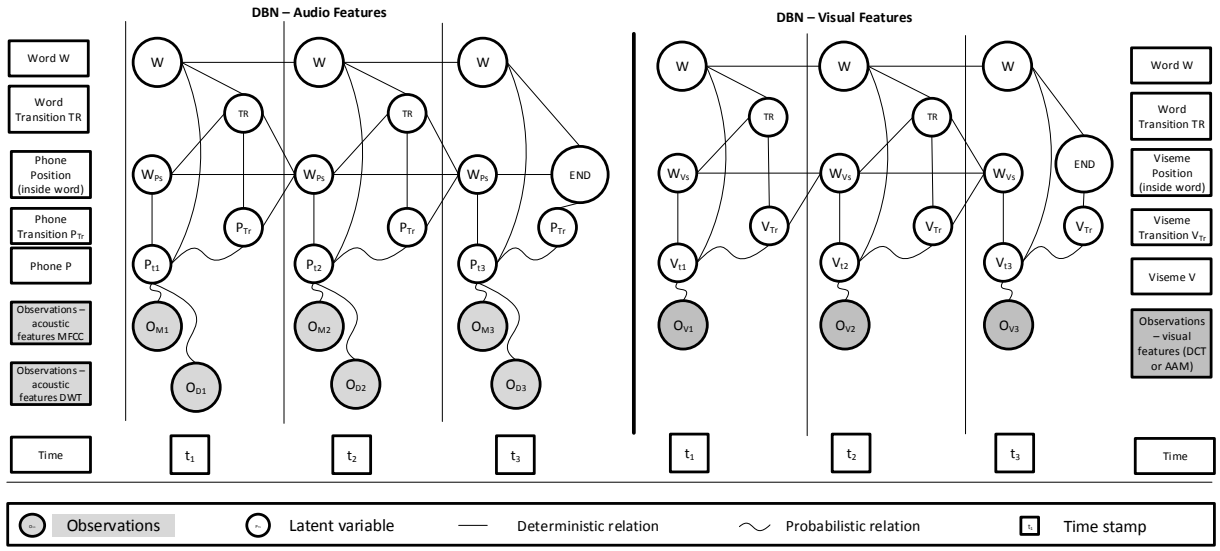


Fig. 1: Scheme of DBN models with two modalities, and asynchrony represented by separate phone / viseme nodes. Figure is split into Audio and Visual parts just for clarificity.

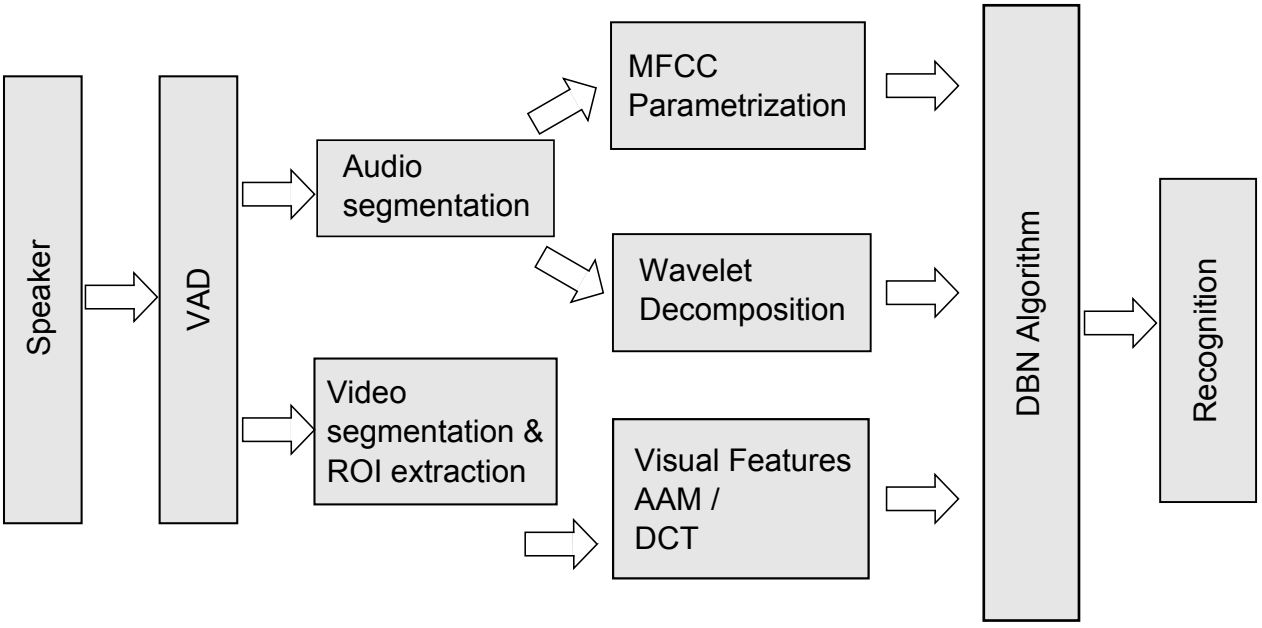


Fig. 2: Data flow in our audio-visual recognition system

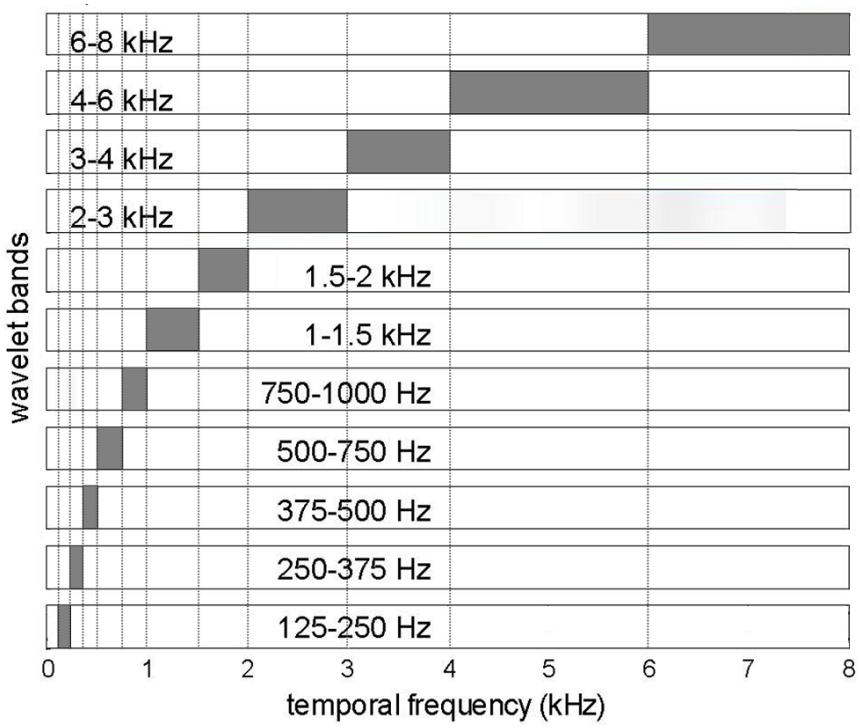


Fig. 3: Discrete wavelet parametrization with optimal decomposition tree. Frequency sub-bands used in audio parametrization. Lowest frequencies (0 - 125 Hz) are discarded from processing, for noise reduction.

### 3.2 Audio Features Extraction

The speech input was first windowed, with Hamming Window of length 20ms and window offset 10ms. For each frame 12 MFCC features and energy feature were extracted. Additionally first and second order derivatives were appended to the MFCC, resulting in a 39 dimensional feature vector $O^M$. Second stream in audio line is for Wavelet parametrization $O^D$. Eleven levels perceptual decomposition procedure with discrete Meyer wavelet decomposition filters were applied to obtain a power spectrum of speech signal. The time discretization for all wavelet sub-bands is unified by summing adequate number of wavelet spectrum samples for high frequencies. Optimal wavelet tree (Fig.3) was found to choose exact boundaries of frequency sub-bands (Gałka and Ziółko 2009). The signal $s$ is decomposed with discrete Meyer high-pass filters g and low-pass h according to the designed perceptual tree. This approach provides decent psycho-acoustic model of the human ear Mel-like frequency characteristics. The parametrization is conducted by measuring different sub-band energy fractions and storing them in a vector of their magnitudes.

### 3.3 Visual Features Extraction

Visual features may be broadly categorized in two approaches (Lan et al. 2010). The first one, top-down approach is derived from a higher-level, model-based features, that utilizes shape and appearance of mouth area images. The second is a bottom-up approach, where model is built from low-level, image-based features which are then compressed using one of dimensionality reduction algorithm, such as discrete cosine transform (DCT) or principal components analysis (PCA). Both type of features require mouth-region extraction executed as an initial step. We are using Viola-Jones algorithm, which is based on an idea of a boosted cascade of weak classifiers, where each one has a high detection ratio and small true reject ratio. Model defining mouth area was described by (Castrillón Santana et al. 2007) and is provided with OpenCV

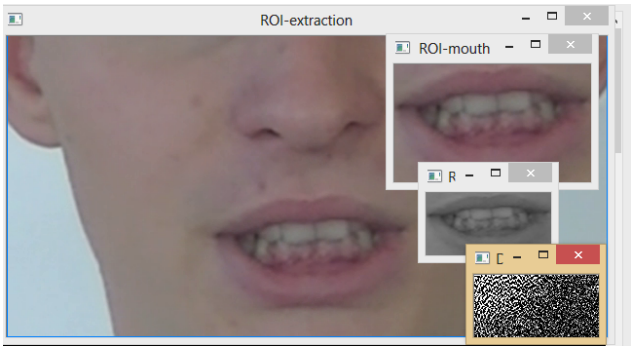


Fig. 4: Mouth region and low-level, image-based features extraction (DCT)

framework. As a result we are gathering rectangle area with varying size, which is later processed. We are testing features from both categories. Visual features are extracted at 25Hz and then interpolated to meet audio features frequency (100Hz).

Image-based features (denoted as **DCT**) are extracted by downsampling ROI to 64x32 pixels and conversion to gray-scale intensity image. Two dimensional discrete cosine transform (DCT) is applied. First 30 coefficients, without DC value and their derivatives, constitute 60-dimensional vector. Following steps are presented in Fig. 4.

Model-based features are using Active Appearance Model (**AAM**) (Cootes et al. 2001) algorithm, that utilizes shape and appearance information. Shape component is formed by concatenating the 2D coordinates of a set of $n$ vertices that are boundary markers. Shape can also be presented in a more compact way, as a mean shape and linear combination of $m$ largest eigenvectors of the covariance matrix. Principal Component Analysis (PCA) is used for finding model parameters. The appearance is defined by pixels that lie inside base shape. Appearance can be represented in similar way to shape component, as a base appearance and linear combination of $k$ appearances (computed by applying PCA to shape normalized training images). Shape and appearance components are concatenated, reweighted (due to their natures: shape are coordinates and appearance are pixel intensities) and processed by final PCA to obtain more compact and decorrelated features. A $z$-score normalization is performed to enhance speaker-variability robustness of visual features.

### 3.4 Database

In this work, we were using the Audiovisual database of polish speech recordings (Igras et al. 2013). It contains three different types of utterances: numbers and commands, 160 short phrases from list of most popular questions asked to Virtual Assistant, using spoken language, and parts of read texts (7 different utterances, like articles, definitions, part of stories). Longer utterances was split to several-words phrases for learning and testing purposes. There are 24 speakers, 13 male and 11 female. Recordings contain only faces, frontal view, on bright background with rather invariant lighting conditions, Full HD quality, 25 frames per second. Each speaker has been recorded for about 10 minutes. Total database length is about 4 hours.

## 4 Experiments

In our testing environment, Audiovisual database was divided into two sets. The training set cosisted of 21 speakers, and 3 remaining speakers were used as the testing set. This was repeated 8 times, to test all speakers, resulting in about 1000 testing utterances. Audio models were additionally trained with more than 10 hours of clean speech samples. In following experiments, we have compared four setups of multi-stream DBN AVASR system with standard HMM-based one, under various SNR conditions. Street noise

was artificially incorporated into testing samples. First setup was standard ASR with HMM and MFCC-only (as described in section 3.2, denoted as *HMM-MFCC* in Table 1) parametrization. Then, we've tested DBN model system, with just single stream, parametrized also with MFCC (*DBN-MFCC*), to see differences in decoding algorithms. In third setup (*DBN-MFCC and DWT*) we tested how multiple parametrization for single modality works under different conditions. Two final experiments gather results for complete system, with both MFCC and DWT features extracted from audio signal, while DCT (*DBN-AV with DCT*) and AAM (*DBN-AV with AAM*) visual features added to DBN decoder accordingly. All results are summarized in Table 1. It shows that incorporating visual modality within DBN-based system may be beneficial to improve phrase recognition rate, especially in noisy condition, regardless of used visual features.

Table 1: Phrase recognition rates for all experiments, under various audio conditions - SNR

| Experiment | 0dB | 5dB | 10dB | clean |
|---|---|---|---|---|
| HMM-MFCC only | 63.69% | 78.97% | 87.50% | 96.73% |
| DBN-MFCC only | 63.79% | 77.98% | 86.21% | 92.86% |
| DBN-MFCC and DWT | 64.68% | 80.95% | 88.69% | 93.45% |
| DBN-AV with DCT | 75.69% | 81.65% | 85.42% | 88.10% |
| DBN-AV with AAM | 76.39% | 81.55% | 85.22% | 87.70% |

## 5 Conclusions

In this paper we have described audio-visual speech recognition system with Dynamic Bayesian Networks used for modeling multiple streams and modalities. On the basis of results collected in Table 1 we can state that the DBN model can be beneficial when SNR is lower, even when only single modality with many parametrizations is taken into account. However, usage of more complicated modeling technique (DBN) may be unjustified when environment does not influence communication between human and computer. This is related to more demanding requirements during DBN model training. These results are consistent with results presented by (Gowdy et al. 2004). Incorporating visual modality, which is not influenced by audio noise is also beneficial, especially in low SNR conditions. We've gathered very similar results for low-level (DCT) and higher-level (AAM) visual features, which is distinct than results presented by (Lan et al. 2010), where authors prefers Active Appearance Model. This is probably due quality of our audio-visual corpus, where invariant lighting conditions results in good recognitions even for simpler (DCT) visual features.

## Acknowledgments

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

## References

Bilmes, J. & Bartels, C. (2005), 'Graphical model architectures for speech recognition', *Signal Processing Magazine, IEEE* **22**(5), 89–100.

Castrillón Santana, M., Déniz Suárez, O., Hernández Tejera, M. & Guerra Artal, C. (2007), 'Encara2: Real-time detection of multiple faces at different resolutions in video streams', *Journal of Visual Communication and Image Representation* pp. 130–140.

Cootes, T., Edwards, G. & Taylor, C. (2001), 'Active appearance models', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **23**(6), 681–685.

Gałka, J. & Ziółko, M. (2009), Wavelet parametrization for speech recognition, *in* 'Proceedings of an ISCA tutorial and research workshop on non-linear speech processing NOLISP 2009, VIC'.

Gowdy, J., Subramanya, A., Bartels, C. & Bilmes, J. (2004), 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–993–6 vol.1.

Igras, M., Ziółko, B. & Jadczyk, T. (2013), 'Audiovisual database of polish speech recordings', *Studia Informatica* **33**(2B), 163–172.

Lan, Y., Theobald, B.-J., Harvey, R., Ong, E.-J. & Bowden, R. (2010), Improving visual features for lip-reading., *in* 'AVSP', pp. 7–3.

Minka, T., Winn, J., Guiver, J., Webster, S., Zaykov, Y., Yangel, B., Spengler, A. & Bronskill, J. (2014), 'Infer.NET 2.6'. Microsoft Research Cambridge. http://research.microsoft.com/infernet.

Murphy, K. P. (2002), Dynamic bayesian networks: representation, inference and learning, PhD thesis, University of California, Berkeley.

Potamianos, G., Neti, C., Gravier, G., Garg, A. & Senior, A. (2003), 'Recent advances in the automatic recognition of audiovisual speech', *Proceedings of the IEEE* **91**(9), 1306–1326.

Saenko, K. & Livescu, K. (2006), An asynchronous dbn for audio-visual speech recognition, *in* 'Spoken Language Technology Workshop, 2006. IEEE', pp. 154–157.

Shivappa, S. T., Trivedi, M. M. & Rao, B. D. (2010), 'Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey', *Proceedings of the IEEE* **98**(10), 1692–1715.

Ziółko, M., Gałka, J., Ziółko, B., Jadczyk, T., Skurzok, D. & Masior, M. (2011), 'Automatic speech recognition system dedicated for polish', *Proceedings of Interspeech, Florence* .

Web-1: http://www.emgu.com/, consulted 01 Apr. 2015.