Viseme-Dependent Weight Optimization for CHMM-Based Audio-Visual Speech Recognition

Alexey Karpov[1], Andrey Ronzhin[1], Konstantin Markov[2] and Miloš Železný[3]

[1] St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences, Russia
[2] Human Interface Laboratory, The University of Aizu, Fukushima, Japan
[3] University of West Bohemia, Pilsen, Czech Republic

{karpov,ronzhin}@iias.spb.su, markov@u-aizu.ac.jp, zelezny@kky.zcu.cz

Abstract

The aim of the present study is to investigate some key challenges of the audio-visual speech recognition technology, such as asynchrony modeling of multimodal speech, estimation of auditory and visual speech significance, as well as stream weight optimization. Our research shows that the use of viseme-dependent significance weights improves the performance of state asynchronous CHMM-based speech recognizer. In addition, for a state synchronous MSHMM-based recognizer, fewer errors can be achieved using stationary time delays of visual data with respect to the corresponding audio signal. Evaluation experiments showed that individual audio-visual stream weights for each viseme-phoneme pair lead to relative reduction of WER by 20%.

Index Terms: multimodal speech, audio-visual processing, Hidden Markov Models, asynchrony, significance weights

1. Introduction

Audio and visual speech signals supplement each other and when combined, they are capable of improving the performance (accuracy and robustness) of automatic speech recognition systems. An open and important problem is the efficient fusion of the information conveyed by each signal. Optimization of weights assigned with every modality allows the system to efficiently adapt to any noisy environment.

There are two well-known basic approaches for information fusion in audio-visual speech recognition (AVSR) systems: feature (early) fusion and decision (late) fusion. In the last decade, both these ways have been comprehensively studied and many methods based on Hidden Markov Models (HMM), which are extension of stochastic chains proposed by the Russian mathematician of St. Petersburg University A.A. Markov in 1900s, or Dynamic Bayesian Networks have been developed: Multi-Stream HMM (MSHMM), Factorial HMM, Coupled HMM (CHMM), etc. Currently, the best results have been achieved by those methods, which allow modeling of natural asynchrony between auditory speech signals and corresponding visually-observed motion of lips and other face mimics. The essence of the bimodal speech asynchrony problem is that the phoneme and viseme flows in natural speech are not completely synchronized in time. It is partially caused by the co-articulation phenomenon in the course of speech production that reveals itself differently on two speech modalities and causes an asynchrony between them. Thus, an efficient information fusion model has to be able to cope with the bimodal speech asynchrony challenge. Recently, some new HMM-based systems able to decode speech in a state asynchronous framework have been developed, for instance, Coupled HMMs [1] or asynchronous Multi-Stream HMMs [2]. In all of these models, the optimal modality weighting is the main challenge.

In the last years, in order to increase the system's robustness many approaches for estimation of audio-visual (AV) stream reliability and related weights were proposed. Some are based on an analysis of acoustic or/and visual conditions (signal-to-noise ratio - SNR) of the speech data [3]. Others use the maximum likelihood criterion with normalization [4]. Dynamic weighting techniques for on-line weight adaptation to changing environment based on entropy-based modality confidence estimators have also been proposed [5]. However, these methods assign significance weight to the whole feature stream, though some data components (models of visemes and phonemes, for example) may have more influence on the recognition performance than others. The significance of each phone or viseme model and individual modality weights for the bimodal speech recognition process has not yet been thoroughly investigated. Preliminary studies on this topic, such as [6] or [7], have found some dependence of weights on speakers, utterances and AV models based on MSHMMs. In this paper, we attempt to fill the gap in this area and propose viseme-dependent modality weights. This approach was implemented in an audio-visual Russian speech recognizer, which is based on state asynchronous 2-stream Coupled Hidden Markov Models.

2. CHMM-based AVSR system

In this section, we describe a late fusion AV Russian speech recognizer that relies on CHMMs, which are transformed to equivalent left-to-right 2-stream HMMs with tied observation densities based on Gaussian mixtures.

2.1. Fusion of speech modalities

Coupled HMM is a collection of parallel HMM, one for each data stream, where the hidden states at time t for each HMM are conditioned by the hidden states at time $t-1$ for all the related HMMs [1]. One channel is provided for the audio stream and another for the video stream. There are two state variables in the joint AV model, and at any time t, the state of the model is determined by these multinomial variables. The advantage of such configuration is that it allows unsynchronized progression of the two chains, while encouraging the two streams to assert temporal influence on each other. The overall dynamics of the AV speech is determined by both the streams at one time.

A simple way to transform the CHMM to an equivalent HMM that keeps all the properties of the former model was proposed in [8]. A similar approach is used in our recognition system as well. Transformed HMM for a AV speech unit contains all the combinations of parallel states of the corresponding CHMM. In the CHMM model, the two streams are independent, and 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 the CHMM-to-HMM conversion of the hidden states. The resulting 2-stream HMM is shown in Figure 1. 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 decoding. The parameters of 2-stream HMMs are obtained by the Baum-Welch (expectation-maximization) algorithm with maximum likelihood estimation using bimodal training data.

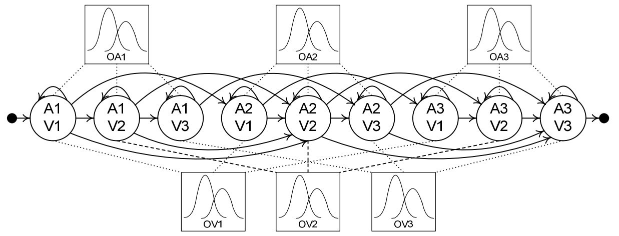

Figure 1: *CHMM to 2-stream HMM transformation.*

2.2. Acoustic and visual speech feature extraction

As acoustic features we used 12-dimentional Mel-Frequency Cepstral Coefficients (MFCC) calculated from 26 channel filter bank analysis of 20 ms long frames with 10 ms overlap. Thus, the frequency of audio feature vectors is 100 Hz. Cepstral Mean Subtraction is applied to audio feature vectors.

The visual (articulatory) features used in the recognizer are based on the work [9] and for their extraction we utilized the source code of the Intel OpenCV-based AVCSR project. The visual parameters are calculated as a result of the following processing steps: multi-scale Haar-based face detection in frames of video data with 25 fps using a boosted cascade classifier; mouth region detection with two cascade classifiers (for mouth and mouth-with-beard) within the lower part of the face; normalization of detected mouth images to 32×32 pixels; mapping to a 32-dimentional feature vector using the principal component analysis (PCA); up-sampling and interpolation of the vectors sequence to 100 Hz in order to correspond with the audio vectors frequency; visual feature mean normalization; concatenation of the consecutive feature vectors into one vector to store the dynamic information in the feature data; viseme-based linear discriminant analysis (LDA). This processing produces 10-dimentional articulatory visual feature vectors with the frequency of 100 Hz. In our previous study [10], this pixel-based visual feature set was compared with the proposed geometry-based visual features describing the shape and configuration of the lips. In those experiments, the word error rates were almost identical for both methods. For this study, we chose the pixel-based parameterization because it requires fewer computations.

2.3. Viseme-dependent stream significance weights

Each CHMM represents one phoneme-viseme pair, and to model audio speech signals, we need more HMMs than for the visual speech modeling only. This is because there are about 10 to 15 visually distinguishable speech units (this quantity is language-dependent) only, while the number of acoustic phonemes is around 42 to 50. According to our previous studies [10], the best recognition results are achieved with 10 visual units (see Table 1).

Table 1. *Viseme classes and phoneme-to-viseme mapping.*

Class	Viseme type	Corresponding phonemes
V1	silence (neutral)	sil (pause)
V2	wide-opened mouth unrounded vowels	a, a!, e! (stressed)
V3	unrounded vowels	e, i, i!, y, y!
V4	rounded vowels	o!, u, u!
V5	labial consonants	b, b’, p, p’, m, m’ (soft)
V6	labio-dental cons.	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

In our system, as joint audio and visual unit models we use 48 CHMMs corresponding to all of the Russian phonemes. After tying the output densities of corresponding viseme models according to the mapping in Table 1 and Figure 1, we got 30 output densities for the visual data stream and 144 untied output densities for the acoustical feature stream.

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 the standard MSHMM- or CHMM-based recognizers, this is made by setting AV 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. It is proposed 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:

$$P(O_t \mid \lambda_{avunit}) = \prod_{s \in \{audio, video\}} P(O_t^s \mid \lambda_{avunit})^{\gamma_{avunit}^s} \qquad (1)$$

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.

3. Evaluation experiments

In this section, we describe the setup and results of the experiments with our Russian AVSR system using different information fusion models, as well as stream- and viseme-dependent importance weights.

3.1. Audio-visual Russian speech corpus

No multimodal corpus is available, which contains segmented recordings of Russian speech. The audio-visual continuous speech database used in this study has been privately recorded in office environment and contains pronunciations of phonetically-balanced sentences uttered by 10 speakers, both men and women. All of them are native Russian speakers with normal articulation in the age of 20 to 70 years old (31 in average). The content was chosen to maximize the statistical coverage of context-dependent Russian phonemes and visemes. The recording session for each speaker lasted about 20-30 minutes; altogether, the speakers uttered 1500 phrases. Sony DCR-PC1000 digital camcorder was used to capture video data with 720x576x25 fps (we are going to use a high-speed video camera with fps ≥ 100 in our future research) and Sony DC-50 microphone located at 15-20 cm from the speaker’s mouth was used for speech sound acquisition. We have developed DirectShow-based software which guaranties the synchronization of the audio and video data streams. The audio data format is: 22 KHz sampling rate, mono, SNR ≈ 25 dB. For training, we selected 60% of each speaker utterances containing phonetically-balanced phrases (90 sentences up to 8 words). The rest of the data consisting of 3 to 6 connected digit long utterances were used for the testing.

3.2. Comparison of audio and video fusion models

Several multimodal and unimodal speech recognizers with different information fusion models have been implemented using the HTK toolkit and compared in terms of recognition accuracy. Babble noise (the so-called "cocktail party" noise) was added to the clean auditory signal with varying SNR in the range from 25 dB to 0 dB. Figure 2 shows the word recognition rates (WRR) for the audio-only recognizer, the video-only recognizer and two multimodal systems based on MSHMMs and CHMMs. These results show the advantage of the multimodal speech recognition systems in comparison with both unimodal recognizers. In AVSR system, we apply global stream weights such that their sum is equal to 2, i.e.: $\gamma_{str}^{video} + \gamma_{str}^{audio} = 2.0$. They are automatically tuned to get the minimal extremum of the WER function in clean speech condition. For the original test data with SNR of 25 dB, the optimal stream significance weights were: $\gamma_{str}^{video} = 0.6; \gamma_{str}^{audio} = 1.4$. It can be seen that CHMM-based system outperformed the MSHMM-based one in all SNR conditions, with absolute improvement in word recognition rate varying from 0.4 to 6.8%. This advantage of the CHMM-based system is explained by its ability to cope with the natural non-stationary asynchrony between the auditory and the visual speech cues (at least, within the model boundaries). However, the standard MSHMM assumes that audio and video observations are synchronous, although allows the audio and video components to have different importance coefficients to the overall observation likelihood.

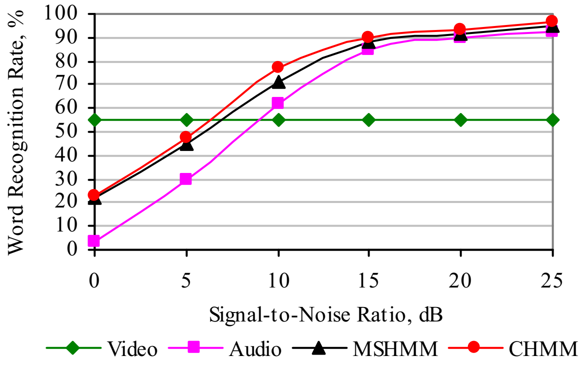

Figure 2: *WRR vs. SNR for four speech recognition models with optimal stream weights.*

It is known from previous studies [11], that visemes usually lead in phoneme-viseme pairs. Moreover, in the beginning parts of phrases, the visual units leave the corresponding phonemes behind more noticeably than in the rest of phrases. Thus, our next experiments were aimed at estimating the influence of the constant audio/video signal shifts (phasing) on the recognition accuracy. Figure 3 presents the results of experiments with the MSHMM-based system, where first the audio data (feature vectors) and then the video data were delayed relatively to the other modality stream by 120/80/40 ms (duration of video frame), respectively. One can notice that the best results were achieved, when the stationary delay of the audio stream was 80 ms (V80A), and a bit lower results - for 40 ms delay (V40A). All other attempts resulted in speech recognition accuracy degradation. These experiments have demonstrated the asynchrony problem between auditory and visual speech features, and that a short shift of the video data can increase the WRR of the state synchronous AVSR system from 94.5 to 96.2%. Nevertheless, this is still worse than the CHMM-based approach, which has reached maximum recognition accuracy for clean speech data of 96.6%. For the CHMMs, signal phasing could not improve the WRR because these models allow state asynchronous decoding and already account for this phenomenon.

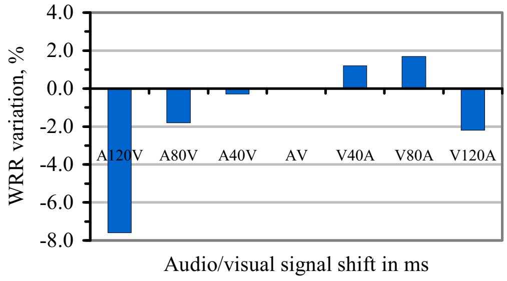

Figure 3: *Dependence of WRR variation on video/audio signal delays in AVSR (clean speech condition).*

3.3. Experiments on viseme-dependent optimization of audio-visual stream weights

Next experiments were focused on the influence of the significance weights on the recognition performance. Firstly, we investigated how the WRR depends on the stream-dependent weights. The results are summarized in Figure 4. Five solid-line curves show WRRs for the following global static AV weight pairs: 1.9:0.1, 1.4:0.6, 1.0:1.0, 0.5:1.5; 0.1:1.9. The dashed-line curve shows the CHMM-based AVSR system with dynamic stream weights, which were adapted to audio data quality by varying the audio weight in the range from 0.0 to 2.0 with a step of 0.1 in order to minimize the WER value (20 weight pairs). Clearly, this curve intersects other curves at the point of best WRR for each SNR.

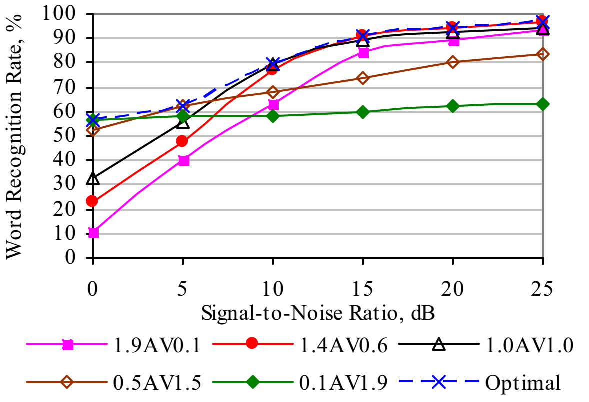

Figure 4: *WRR vs. SNR for the CHMM-based recognizer with five static and one dynamic global stream weights.*

Previous experiments presented the results for the stream-dependent weights only, i.e. for the case when one significance weight is set for the audio stream and another for the visual one. Below, we investigate the contribution of every visemic-phonemic HMM-based model and its significance by assigning individual weights to the CHMM streams. In these experiments, the optimal stream weights were determined by consecutively changing the weight for each of the 10 viseme classes (see Table 1) with a discrete step of 0.1 in the range ± 0.4 from the best dynamic stream weight γ_{str}^{video} in order to minimize the WER. Table 2 presents the best WRRs for both viseme-dependent and stream-dependent weights (the baseline AVSR system). As the results show, by optimizing the viseme-dependent weights it is possible to achieve better WRR especially for low SNRs.

Table 2. *The best word recognition rates for the stream- and viseme-dependent significance weights.*

Model type \ SNR	Word Recognition Rate, %		
	5 dB	10 dB	25 dB
Stream-dependent weights (baseline)	62.6 (0.5 : 1.5)	79.4 (1.0 : 1.0)	96.6 (1.4 : 0.6)
Viseme-dependent weights (proposed)	65.0	80.4	97.3

Estimations of the best viseme-dependent weights for each visual unit (see Table 1) depending on the SNR are given in Figure 5. Dashed lines denote the optimal global video stream weights for each SNR. Some common tendencies can be observed after analysis of this chart:

- At low SNR increasing importance of video data is efficient for rounded vowels, labial and labio-dental consonants, and wide-opened mouth vowels, i.e.:

$$\gamma_{V2,4,5,6}^{video} > \gamma_{str}^{video}, if\ SNR < 10dB \qquad (2)$$

- On the contrary, at high SNR it is better to decrease the video stream weights for all the visemes from the previous statement as well as for the alveolar fricatives, i.e:

$$\gamma_{V2,4,5,6,7}^{video} < \gamma_{str}^{video}, if\ SNR > 20dB \qquad (3)$$

- The video stream weight for the silence model should be reduced in any conditions: $\gamma_{V1}^{video} < \gamma_{str}^{video}$. This could be explained by the fact that the model corresponding to the pause is ambiguous: some human beings start speaking with the closed mouth, but others – with slightly or fully opened mouth as a neutral lips position. The silence in the acoustical sense is more or less invariant.

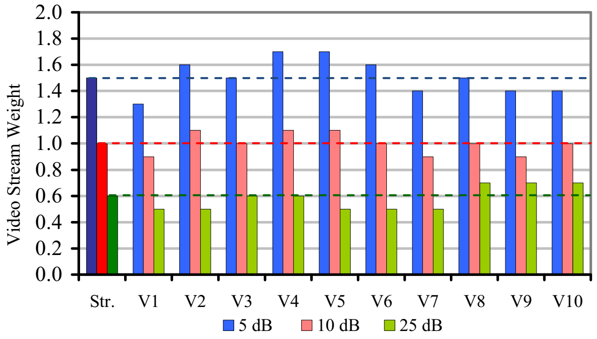

Figure 5: *Optimal viseme-dependent weights of visual speech modality for SNR = 5/10/25 dB.*

The experiments with viseme-dependent weights show that the visemes, which exhibit significant lips motion have the biggest influence on the speech recognition accuracy. Besides, in noisy environment, these visemes contribute more than others, and it is easier to recognize them, because their models are far from "the average model". Their negative contribution in clean speech condition is explained by the weakness of the visual features in comparison to the acoustic features, which can better distinguish between all the vowels and fricatives. In conditions with medium SNR (≈ 10 dB), viseme-dependent weights and word error rates are not much different from the optimal global stream weights.

4. Conclusions

In this paper, we presented experiments with viseme- and phoneme-dependent significance weights in an AVSR system based on both Coupled and Multi-Stream HMMs. The results of these experiments demonstrated the positive effect of using independent stream weights for each AV model, sustaining the hypothesis that setting the AV weights must to be done on a per-viseme basis, because some visemes are more important than others in high audio-noise environment. A number of common tendencies to discriminate weights for the visual classes and their influence on the AVSR accuracy were explored. The application of individual AV stream weights for each AV speech unit model provided relative WER reduction by 10-20% depending on acoustical environment.

5. Acknowledgements

This research is supported by the Grant of the President of Russia (project MK-64898.2010.8), the Ministry of Education and Science of Russia (contract No. П2579), and the Ministry of Education of the Czech Republic (project No. ME08106).

6. References

[1] Nefian, A.V., Liang, L.H., Pi, X., Xiaoxiang, X., Mao, C. and Murphy, K., "A coupled hmm for audio-visual speech recognition", In Proc. ICASSP-2002, Orlando, USA, 2002.

[2] Neti, C., Potamianos, G., Luettin, J., et al., "Audio-visual speech recognition", In Final Workshop 2000 Report, Baltimore, USA, 2000.

[3] Heckmann, M., Berthommier, F. and Kroschel, K., "Noise adaptive stream weighting in audio-visual speech recognition", EURASIP Journal on Applied Signal Processing, 1:1260-1273, 2002.

[4] Tamura, S., Iwano, K. and Furui, S., "A stream-weight optimization method for audio-visual speech recognition using multi-stream HMMs", In Proc. ICASSP-2004, Montreal, Canada, 2004.

[5] Gurban, M., Thiran, J.-P., Drugman, T. and Dutoit, T., "Dynamic modality weighting for multi-stream HMMs in audio-visual speech recognition", In Proc. ICMI-2008, Chania, Greece, pp. 237-240, 2008.

[6] Glotin, H., Vergyri, D., Neti, C., Potamianos, G. and Lüttin, J., "Weighting schemes for audio-visual fusion in speech recognition", In Proc. ICASSP-2001, Salt Lake City, Utah, USA, pp. 173-176, 2001.

[7] Patel, P. and Ouazzane, K., "Comparison of fixed and variable weight approaches for viseme classification", In Proc. IASTED International Conference on Signal and Image Processing SIP-2007, Honolulu, USA, pp. 110-115, 2007.

[8] Chu, S.M. and Huang, T.S., "Multi-Modal sensory Fusion with Application to Audio-Visual Speech Recognition", In Proc. Multi-Modal Speech Recognition Workshop-2002, Greensboro, USA, 2002.

[9] Liang, L., Liu, X., Zhao, Y., Pi, X. and Nefian, A., "Speaker independent audio-visual continuous speech recognition", In Proc. International Conference on Multimedia and Expo ICME-2002, Lausanne, Switzerland, 2002.

[10] Cisar, P., Zelinka, J., Železný, M., Karpov, A. and Ronzhin, A., "Audio-Visual Speech Recognition for Slavonic Languages (Czech and Russian)", In Proc. 11-th International Conference on Speech and Computer SPECOM-2006, St. Petersburg, Russia, pp. 493-498, 2006.

[11] Karpov, A., Tsirulnik, L., Krnoul, Z., Ronzhin, A., Lobanov, B. and Železný M., "Audio-Visual Speech Asynchrony Modeling in a Talking Head", In Proc. Interspeech-2009, Brighton, UK, pp. 2911-2914, 2009.