BIMODAL AUDIO-VISUAL AUTOMATIC SPEECH RECOGNITION: AN OVERVIEW

Benoît Lasjaunias

Grenoble Institute of Technology - Phelma
France
bvpla@kth.se

Joaquín Antón Guirao

Polytechnic University of Valencia
Spain
joaag@kth.se

ABSTRACT

Bimodal Audio-Visual Automatic Speech Recognizers (AVASR) happen to be highly relevant in noisy environments. However, the choice of the visual features and the way of combining them with the acoustic features has been a matter of research in the last decades. This paper makes an overview of the latest work in that field. The most used visual features are mainly based on the lip geometry, including sometimes their motion and texture. When combining the two modalities, two different approaches are adopted: feature fusion and decision fusion. The feature fusion combines both audio and visual features in a single feature vector while a single classifier is used to make a decision. On the other hand, the decision fusion works separately with visual and audio modalities by using two classifiers. Then, the final decision is made by combining decisions from both modalities. We have seen that combining methods based on empirical reliability measurements leads to more suitable decisions. Eventually, this work shows that AVASR systems rise the accuracy comparing to the audio-only ASR systems, especially in extreme conditions.

1. INTRODUCTION

Automatic Speech Recognition (ASR) has been increasingly improved during the last decades, due to its particular interest in applications in human-computer interaction. Most of the research and the most successful works have been focused on acoustic-only based recognition. However, the low accuracy of these recognizers in noisy environments leads to the development of multimodal systems. Figure 1 shows the main processing blocks of the

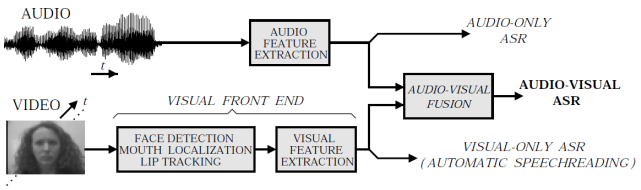

Figure 1: Main processing blocks involved in the AVASR (Source: [1]).

AVSR. Video and audio signals are processed separately for feature extraction for audio-visual fusion. Thus, Audio-Visual ASR (AVASR) mimics the natural-human lip-reading mechanism to improve their speech understanding skills. In the present paper we make an overview of the AVASR related work, focusing on the most recent systems found in the literature. In the Section 2 we summarize the different visual-feature extraction methods used in AVASR. In Section 3, the most-used ways of modality-fusion are presented. Section 4 explicitely describes both archictecture and performance obtained by the different works reviewed. Finally, Section 5 states the conclusions of this state-of-art analysis.

2. VISUAL FEATURES

The improvement in the word error rate due to the combination of vision and audio in automatic speech recognition has been proved in many papers. The question is how to choose and extract relevant vision features for a maximum improvement. Three types of features mainly appear in the literature: the lip-geometry, the lip-motion, and the lip-texture. Some systems combine several of these types, but it is more relevant to study them separately before combining, as done in [2].

2.1. The lip-geometry

By assuming that the speech information is contained into the outline of the lips, one should be able to detect them precisely. Every single system has its own lip detection algorithm, which is often a combination of a face detector and a lip-outline detector, as shown in Figure 2.

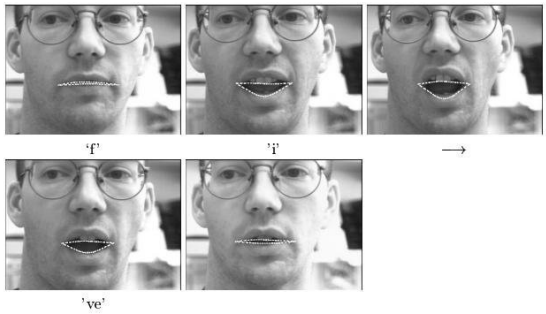

Figure 2: Use of valley for tracking the lip-outline. Word "five" is used on the example by showing snapshots every 60 ms [3].

There are two main categories of information in the lip geometry: the shape, i.e. the outline, and the geometric characteristics, i.e. height, width, area, etc. The outline is often modeled with polynomial coefficients. In [3] the 2D outline of the lips is parameterized by quadratic B-splines, which is a combination of low-degree polynomials, since it requires fewer parameters than a regular polynomial interpolation. Moreover, in order to get the

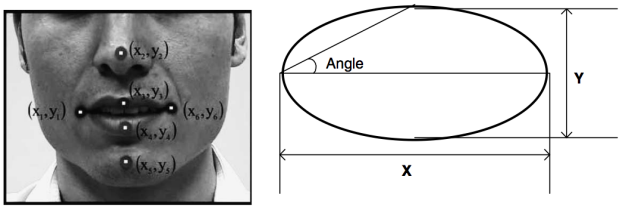

Figure 3: Geometric features used in [2]. Six facial points are used to compute the features: outer-lip horizontal width (X), outer-lip vertical width (Y), outer-lip area (area inside of the ellipse), and angle of the outer-lip corner (angle)

most relevant features, it is possible to find a reduced basis to represent the control points of the outlines by projecting on the affine basis or by applying a Principal Component Analysis (PCA). The affine basis is however insufficient to model all the relevant deformations of the lips. Therefore the method used in [3] is PCA and the authors have come up with the result that 99% of the deformation information of the lip contour lays in the first six components. It is also possible to use both a frontal and a profile vision of the face. Both give different features, the profile view features are the high-contrast edges whereas the frontal view features are the combination of the edges and the intensity valleys, which are extremely relevant in speech recognition.

As mentioned earlier, it is also possible to use the geometrical information of the lips: the width X, the height Y, the area, and the angle. It might be relevant to take into account the first derivative of these parameters, since the motion of the lips is also interesting. In [2], after having tested independently every single parameter, the authors have come up with the three best individual feautures: Y, dY/dt and $d(angle)/dt$, see Figure 3.

The test of several combinations of parameters provides the idea of what the best combination is: $[X, Y, d(angle)/dt]$, as shown in the Table 1. Other researchers present in [4] the articulatory feature sets. Those are close to the geometrical ones but present a higher level of interpretation. The features are: Lip Opening (with values *closed*, *narrow*, *medium*, *wide*), Lip Rounding (*Yes*, *No*), Labio-Dental (*Yes*, *No*), and another one involving the teeth.

2.2. The lip-motion

This feature supposes that important information of the speech is contained in the lips motion. Moreover, the lip tracking processing time must be significantly low, especially in cases where the ASR is used for real-time applications. In [3] the B-splines are also used for the lip velocity: the motion is represented by the coordinates of the splines varying over time. The experience shows that the translation on the horizontal axis is not relevant, because it is due only to the head displacement in the image. Again, by applying a PCA to these features without the horizontal displacement, the obtained feature needs only the six first components to recover 99% of the lip motion information. Researchers in [5] have also used an image-based method to represent the lip motion information. The feature is based on the motion vectors: two matrices V_x and V_y are computed, which are the x and y-coordinates of the motion vectors between two consecutive frames. The final feature set is the 50 first coefficients of the 2D-DCTs, i.e. a vector of length 100. The 2D-DCT presents the advantage to concentrate the energy of the motion vectors matrix into the first coefficient, leading to a sorting of the components. This method takes therefore the 50 most relevant coefficients of the matrices.

2.3. The lip-texture

This third type of feature is based on the common practice of working with intensity information of the lip image. The 2D-DCT transform based on the intensity is computed and the discrimination content is found in it. Then, the feature set is composed of the 50 most discriminative DCT coefficients.

2.4. The Active Appearence Models

The AAMs are not part of any of the upper categories because they are based on the Active Shape Model (ASM) to match the shape, but add also information about the texture. AAM analyze the shape variability of whole faces, represented by landmark points in a low-dimensional space. The important relative difference to image-based transform is that AMMs explicitely capture separately the shape and the texture variation of the face.

3. COMBINING AUDIO AND VISUAL FEATURES IN A SINGLE CLASSIFIER

Models for audio-visual integration can be divided in two main approaches:

1. Early Integration (feature fusion): Visual and acoustic features are combined to create a single feature vector and thus a single recognizer is used.
2. Late Integration (decision fusion): Each stream is processed by an independent classifiers that gives its own output. Later, the output of both visual and acoustic classifier are combined to set the final decision.

In the following sections we are presenting some of the most relevant methods used lately in the audio-visual fusion task.

3.1. Dynamic Bayesian Networks and HMM

Most of the audio-visual models used in AVASR systems including HMMs and its variations are particular cases of the Dynamic Bayesian Networks (DBN). DBNs are direct graphical models of stochastic processes in which the hidden states are represented by individual variables or factors [6].

The single-stream HMM approach is based on the concatenation of both visual and acoustic features in a single feature vector and the use of a HMM. However, the dimension of this vector can be large, causing problems due to the curse of dimensionality. Moreover, this method is not considered the best solution since it can not easily represent the loose timing synchronicity between audio and visual features.

A more complex HMM model is needed when trying to handle a problem with additional complexity for audio-visual correlation and loose synchronicity between sequences. There are several HMM-based approaches that solve this issue including factorial HMM, coupled HMM and multi-stream HMM.

The factorial HMM model (FHMM) for AVASR has two streams and a finite number of states. Thus, the hidden states of each modality contribute to the emission of a single observation. In other words, the hidden state is distributed. The factors (hidden variables) of each stream are independent but both contribute to a single observation.

Evaluation of HMMs for combined visual features

Combined visual feature analysis—speaker independent							
N	4	4	4	4	5	5	5
M	4	8	16	32	8	16	32
Visual feature	Average recognition rate (%)						
X–Y	55.4	57.9	62.5	63.3	63.3	62.9	62.9
Y–d(Y)/dt	47.5	46.7	47.1	42.9	46.7	51.3	52.1
X–Y–d(Y)/dt	67.1	65.0	70.8	70.8	70.4	72.5	70.8
X–Y–angle	50.1	51.7	50.4	51.7	52.5	50.8	50.8
X–Y–d(angle)/dt	63.3	68.3	72.1	72.5	67.9	73.3	74.6
X–Y–d(Y)/dt–d(Y/X)/dt	55.4	60.8	60.0	59.2	60.0	62.5	60.8
X–Y–angle–d(angle)/dt	60.4	60.0	60.8	62.1	59.2	60.0	60.8

N is the number of states and M the number of Gaussian mixture components.

Table 1: Performance of the different single-feature combination evaluated in [2]. They show that the combination $[X, Y, d(angle)/dt]$ is the best to increase the AVASR accuracy

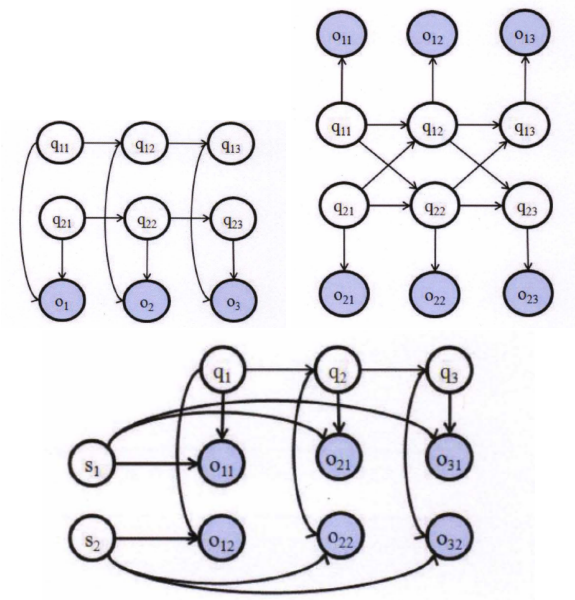

Figure 4: FHMM (up-left), CHMM (up-right), MSHMM (down)

On the other hand, the coupled HMM model (CHMM) [7] allows the hidden nodes from each stream to interact with each other while each of them has its own separate observation. Thus, the CHMM can be seen as a set of HMMs, one for each data stream where all the nodes at time t are conditioned by the nodes at time $t-1$ for all the related HMMs.

Both FHMM and CHMM allow asynchrony between the sequences since different streams have separated hidden state sequences.

On the other hand, the state-synchronous Multi-Stream HMM (MSHMM) structure handles multiple streams for temporal data. In principle, it is used when the streams are synchronous and independent, by assuming that there is a single hidden state for each time slot. Thus, a single hidden state is tied to independent observations for each modality, that are generated by different models. However, there are different variations of the MSHMM that allow to work with non synchronous multi-stream data, which is the common case for AVASR.

3.2. Reliability-based fusion methods

The work presented in [5] solves the multimodal integration by a method called reliability weighted summation (RWS) based on an weighted average of a set of scores. The total log-likelihood is defined by

$$\rho(\lambda_r) = \sum_{n=1}^{N} \omega_n \rho_n(\lambda_r) \tag{1}$$

where ω_n is the weight coefficient for modality n, with $\sum_n \omega_n = 1$, and $\rho_n(\lambda_r)$ are the individual likelihoods (scores) for each modality. Thus, the reliability estimation techniques are applied to estimate the weight coefficients. Ideally when the acoustic speech is noise-free, the difference between the outputs of the HMMs is large. On the other hand, when the signal is noisy the difference becomes small. Thus, the reliability (S_n) used in [5] is defined for the modality n as the difference between the likelihood ratios from the two best class candidates,

$$S_n = \rho_n(\lambda_*) - \rho_n(\lambda_{**}) \tag{2}$$

where $\rho_n(\lambda_*)$ and $\rho_n(\lambda_{**})$ indicate the score for the first and second best class candidate, respectively. Another approach is used in [8] to estimate the reliability

$$S_n = \frac{1}{N_C - 1} \sum_{i=1}^{N_C} (\rho_n(\lambda_*) - \rho_n(\lambda_i)) \tag{3}$$

In practice, a normalized version of the reliabilities are used to estimate the weight factors

$$\omega_n = \frac{S_n}{\sum_j S_j} \tag{4}$$

In [9] a Neural Network (NN) based architecture is presented for combining the decision for both audio and visual measured reliabilities by an adaptive optimal weighting that enables the bimodal recognition system to be robust against different noisy conditions. After the acoustic and visual systems perform recognition separately, their outputs are combined as a weighted sum (see Equation 1). These weights coefficients, involved in the score of the final classifiers, are obtained by a Neural Network that maps optimally the input/output relationship between the two reliabilities and the integration weight. The type of NN used in the proposed method is multilayer perceptrons (MLPs).

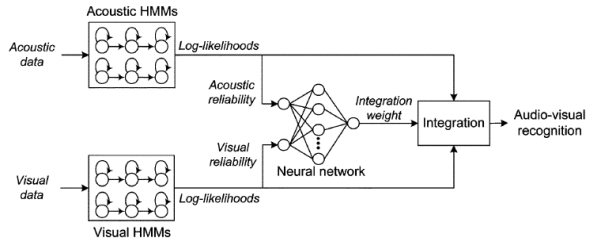

Figure 5: Neural Network based system for integrating Audio-Visual decisions proposed in [9].

In [10] a novel framework based on multimodal fusion by uncertainty compensation is introduced. There, both visual and acoustic features are taken into account with different emphasis depending on their level of uncertainty. Thus, the adaptive fusion rules weight with lower coefficients the degraded features (e.g. noisy acoustic signal, occlused face) by measuring empirically the feature uncertainty, relying thus the decision on the cleaner stream. Moreover, this technique is applicable to either synchronous or asynchronous multimodal architectures (e.g. FHMM, CHMM, MSHMM).

4. REAL SYSTEMS: ARCHITECTURE AND PERFORMANCE

In this section we make a small overview of some of the latest Audio-Visual ASR found in the bibliography.

4.1. Kaucic et al. 1996

In [3] a real-time lip tracker algorithm based on a Kalman filter for detecting the lips contour is presented. The visual information is added to the acoustic features to enhance the performance. Composite features are created by the concatenation of both acoustic and visual features. The authors test the system on a small isolated-word vocabulary for acoustic-only, visual-only and composite features by using Dynamic Time Warping (DTW) as recognition algorithm. Experiments showed that the error rate decreases around the half when combining both visual and acoustic features rather than using acoustic-only features.

4.2. Luettin et al. 1998

In [11] is presented a multi-stream HMM system taking the shape and the intensity of the lips, i.e. geometry and texture features, as visual features. They lower the WER from 3.4% with audio only to 2.6% by combining audio-visual features.

4.3. Nefian et al. 2002

In [6] both CHMM and FHMM models are tested for the audio-visual speech recognition task. Both models allow asynchrony between modalities while preserving their natural correlation over time. For the visual features they use different combinations (1D DCT, 2D DCT, LDA, template) obtaining the best performance for the combination [Window, 2D DCT, LDA]. They make experiments for a speaker dependent isolated word ASR, obtaining the best performance results for the CHMM (see Figure 6). The results

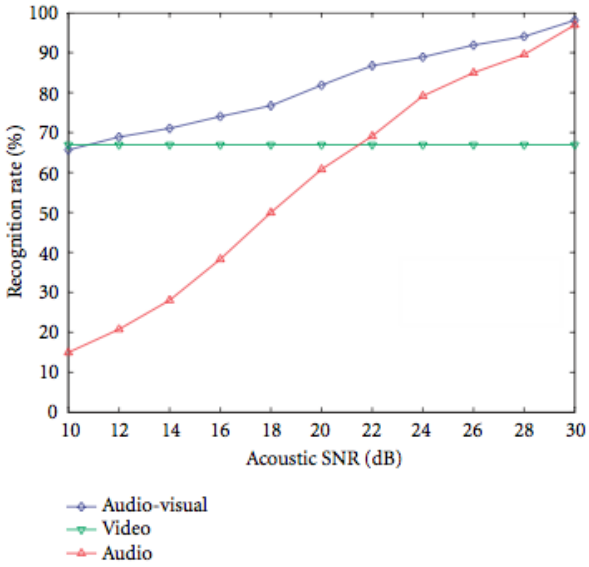

Figure 6: Results presented on [6] for acoustic-only, visual-only, and combined features under different level of signal-to-noise ratio (SNR). The results show a high increase in performance for using combined audio-visual features, especially for low SNRs using the CHMM

show a significant increase on the performance when using audio-visual features compared to only-acoustic features. This gain is especially high for low SNR.

4.4. Kaynak et al. 2004

The system presented in [2] uses geometric visual features based on the lips to enhance the performance of the ASR in noisy enviroments. Their geometric approach is based on 6 facial points to compute the different features (see Figure 3): outer-lip horizontal width (X), outer-lip vertical width (Y), outer-lip area (area inside of the ellipse), and angle of the outer-lip corner (angle). The relevance information of each possible single visual feature is used to find the best configuration for bimodal recognition. They found that the lip apertures and the first derivative of the lip corner angle are the most representative. For the recognition algorithm they use the well-known HMM. Tests are done for acoustic, visual, and combined features under different levels of noise. The results show a highly improved accuracy (around 20%) for using 3 labial geometric features (for $SNR(dB) = 0$). More complete results are shown in Figure 7.

4.5. Çetingül et al. 2006

In [5] a multimodal speaker/speech recognizer is introduced. The acoustic features are represented by the MFCCs, while lip texture and motion is estimated for the visual features. For the lip texture they use the 2D-DCT coefficients of the grey levels. The lip motion is computed by a discriminative analysis of the dense motion vectors. The combination of the audio-visual information is done in the decision layer by using RWS decision rule. The conclusions state that including lip motion modality increases the performance

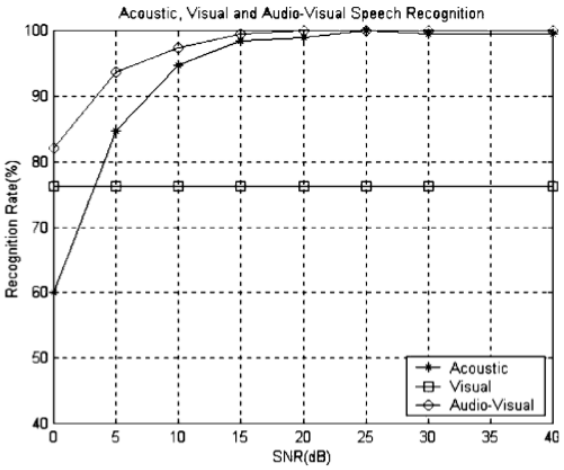

Figure 7: Results presented on [2] for acoustic-only, visual-only, and combined features under different level of signal-to-noise ratio (SNR). The results show a high increase in performance for using combined audio-visual feautures, specially for low SNRs

of the recognizer. Nevertheless, the performance gain achieved is not significant.

4.6. Lee et al. 2008

In [9] the system is divided into three subsystems: visual feature extraction, audio feature extraction and audio-visual fusion, which are optimized independently. HMMs are used to model the visual data, and are improved by the Hybrid Simulated Annealing (HSA) algorithm, which allows escaping from local extrema and reaching the global minimum. The acoustic data correlation is modeled by HMMs with a Gaussian mixture model. The data fusion is computed by an artificial neural network. They obtain a result shown Figure 8, for different noise models and levels. The improvement compared to an audio-only system for $SNR(dB) = 0$ is more than 20%.

4.7. Saenko et al. 2009

In [4], an AVASR using dynamic Bayesian networks (DBN) is introduced. For the visual features they use articulary features such as lip opening or lip rounding. A multi-stream of discriminative articulary features SVM classifiers feeds the input of the DBN, that represents streams allowing asynchrony between them. Experiments show that using articulatory features outperform conventional acoustic-only model's performance by around a 10%.

4.8. Papandreou et al. 2009

In [10] the uncertainty of each feature is exploited to lead to a highly adaptive fusion rule. The recognition is computed with HMM algorithm. As shown Figure 9 the W-label leads to much better results, and means that the weights are optimally updated regarding the uncertainty of the features. The type of features used here is also relevant to notice: the visual features are extracted with the AAM framework, which contains an optimal number of

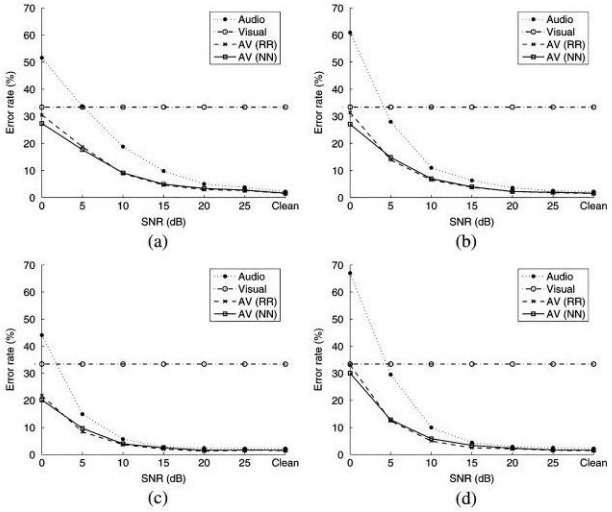

Figure 8: Performance of the acoustic-only, visual-only and composite features for the AV digit-ASR presented in [9] for different types of acoustic noise and levels of SNR(dB): (a) WHT, (b) F16, (c) FAC, (d) OPR.

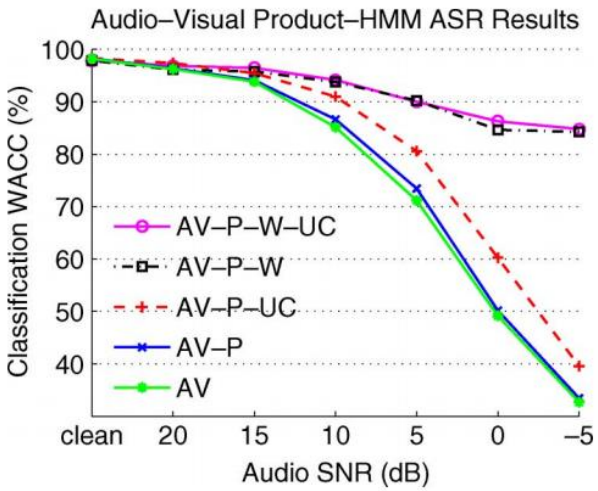

Figure 9: Performance of the Product-HMM-based ASR presented in [10]. The authors show that the W-label updating system regarding the uncertainty of the features leads to much better accuracy.

coefficients for the shape and the texture of the lips: six-shape/six-texture visemic AAM visual feature set. They reach a classification accuracy of around 85% for a SNR of $0dB$.

5. CONCLUSIONS

In this work we made an overview of the latest and most relevant researches on the field of Audio-Visual ASR. Both feature extraction methods and audio-visual fusion mechanisms were analyzed. Generally, geometric information of the lip-outline and/or its motion is commonly used for enhancing the accuracy as visual features. For that purpose, different feature extractors have been introduced in the literature. On the other hand, there are different approaches for combining both modalities, either by using feature-fusion or decision-fusion approaches. Some systems introduce a mechanism to weight the uncertainty of each modality regarding empirical measurements in order to make a more consistent decision. It is shown that this approach leads to better performance. Nevertheless, it is seen that the improvement of adding the visual modality is moderate but it becomes significantly high for low signal-to-noise rates. However, the comparison between systems is not a trivial issue since different methods are tested for different purposes and with different databases (single-word, digit, etc).

6. REFERENCES

[1] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. *Audio-Visual Automatic Speech Recognition : An Overview*. MIT Press, 2004.

[2] M. Kaynak. Lip geometric features for human-computer interaction using bimodal speech recognition: comparison and analysis. *Speech Communication*, 43(1-2):1–16, June 2004.

[3] Robert Kaucic, Barney Dalton, and Andrew Blake. Real-time lip tracking for audio-visual speech recognition applications. In *Proc. European Conference on Computer Vision, volume II of Lecture Notes in Computer Science*, pages 376–387, 1996.

[4] Kate Saenko, Karen Livescu, James Glass, and Trevor Darrell. Multistream articulatory feature-based models for visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1700–1707, 2009.

[5] H. E. Çetingül, E. Erzin, Y. Yemez, and A. M. Tekalp. Multimodal speaker/speech recognition using lip motion, lip texture and audio. *Signal Process.*, 86(12):3549–3558, December 2006.

[6] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP J. Appl. Signal Process.*, 2002(1):1274–1288, January 2002.

[7] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, CVPR '97, pages 994–, Washington, DC, USA, 1997. IEEE Computer Society.

[8] Trent W. Lewis and David M. W. Powers. Sensor fusion weighting measures in audio-visual speech recognition. In *Proceedings of the 27th Australasian conference on Computer science - Volume 26*, ACSC '04, pages 305–314, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[9] J. S. Lee and C. H. Park. Robust Audio-Visual Speech Recognition Based on Late Integration. *Multimedia, IEEE Transactions on*, 10(5):767–779, 2008.

[10] George Papandreou, Athanassios Katsamanis, Vassilis Pitsikalis, and Petros Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 17(3):423–435, March 2009.

[11] J. Luettin and S. Dupont. Continuous audio-visual speech recognition. In *Proc. 5th European Conference on Computer Vision*, volume II of *Lecture Notes in Computer Science*, pages 657–673. Springer Verlag, 1998.