Noise Suppression in Tele-Lectures using Bi-Modal Feature Extraction

[1]E.S.Selvakumar, [2]S.Shanmuga Priya

[1]Department of Information Technology,
Periyar Maniammai University, Vallam, Thanjavur-613 403.

[2] Department of Computer Science and Engineering
St.Joseph College of Engineering and Technology,
Elupatti, Thanjavur.

Abstract

Automatic Speech Recognition (ASR) is an essential component in many Human-Computer Interaction systems. A variety of applications in the field of ASR have reached high performance levels but only for condition-controlled environments. In this project, we reduce the noise in the video lectures using bi-modal feature extraction. Audio signal features need to be enhanced with additional sources of complementary information to overcome problems due to large amounts of acoustic noise. Visual Information extracted from speaker's mouth region seems to be promising and appropriate for giving audio-only recognition a boost. Lip/Mouth detection and tracking combined with traditional Image Processing methods may offer a variety of solutions for the construction of the visual front-end schema. Furthermore, Audio and Visual stream fusion appears to be even more challenging and crucial for designing an efficient AV Recognizer. In this project, we investigate some problems in the field of Audio-Visual Automatic Speech Recognition (AV-ASR) concerning visual feature extraction and audio-visual integration to reduce noise in the video lectures.

Keywords: *ASR, Audio-visual automatic speech recognition, Feature extraction, Multi-stream HMM.*

1. Introduction

In recent years, the field of Audio Visual Speech Recognition (AVSR) has proved to be of significant interest for many researches, as the traditional Audio Automatic Speech recognition (ASR) systems seem to work only for relatively controlled environments. A major problem of ASR is robustness under channel and environmental noise.

Many techniques have been investigated to improve the recognition under noisy conditions, including mainly enhancement of the audio signal, applying noise resistant parameterization, and identifying speech in those sub-bands of the spectrum that the speech signal is dominant. However, for ASR to approach human levels of performance and for speech to become a truly pervasive user interface, we need novel, nontraditional approaches that have the potential of yielding dramatic ASR improvement and to reduce the noise in the video lectures. Lip reading, as an alternative source of information, consist such a different approach which is not affected by the acoustic environment and noise, and it possibly contains the greatest amount of complementary information to the acoustic signal.

AV-ASR systems can outperform audio-only recognizers, particularly in environments where background noises and multiple speakers exist. We combine audio and visual information in deciding what has been spoken, especially in noisy environments. One of the key properties of bimodal speech to emerge from such analysis is that of complementarity: Features that are the hardest to distinguish acoustically are the easiest to distinguish visually, and vice versa.

The sensory integration of auditory and visual information in speech perception and the complementarity between these modalities shows clearly in experiments that independently vary auditory and visual information.

Audio-visual fusion is an instance of the general classifier combination problem. In our case, two observation streams are available (audio and visual modalities) and provide information about hidden class labels, such as HMM states, or, at a higher level, word sequences. Each observation stream can be used alone to train single-modality statistical classifiers to recognize such classes.

One of the main challenges in AV-ASR systems is the audio-visual information integration problem. The main issues in information integration are, (a) the class conditional dependence assumption made across streams, (b) the

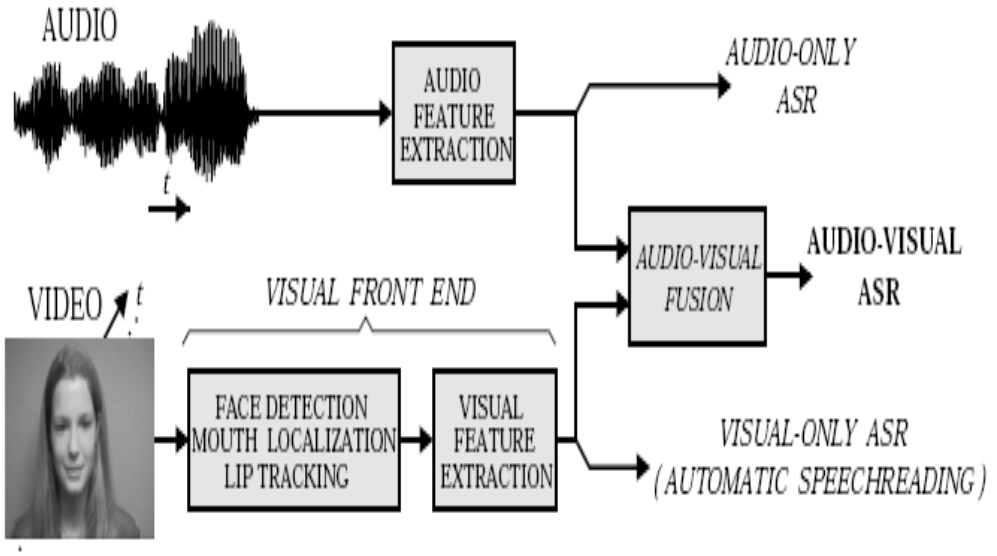

Figure 1.1: The main concept of an AV-ASR system

level (e.g. frame, phone, word) of integration, and (c) the kind (e.g. feature, partial likelihood, partial decision) of integration. Generally, speaking it is widely accepted that there are two kinds of integration: feature fusion and decision fusion.

Our work reported in this paper is summarized below:

Chapter 2 represents the visual front-end method we used to extract the visual features, including some improvements we made in a state of the art visual front-end, such as rotation correction and scaling normalization of the obtained Region of Interest (ROI).

Chapter 3 enumerates methods for computing stream confidence and estimating stream weights including the K-means based method we applied in our weight estimation schema.

Chapter 4 represents the noise filtering in the video stream using the Butterworth filter experimental results.

2. Feature Extraction

Alternatively, lip contour geometric features are used, such as mouth height and width.

Here we have a three type of feature extraction algorithms.

2.1. Overview

A main problem in the research field of audio as well as audio-visual ASR is feature extraction. If the extracted features from the audio signal or the speaker's images are carefully chosen, it is expected that the feature set will extract the relevant information in order to be performed an efficient recognition, classification. In the case of audio, mel-frequency cepstrum coefficients (MFCC) are often used in applications such as speech recognition, voice recognition, audio compression and music information retrieval.

Various sets of visual features have been proposed over the last 20 years for this purpose. In general, they can be grouped into three categories: High level lip contour (óhape) based features; low level appearance (pixel) based ones and finally a combination of both. In the first approach, the speaker's inner (and/or outer) lip contours are extracted from the image sequence. A parametric or statistical lip contour model is then obtained and the model parameters are used as visual features.

2.1.1 Lip geometry estimation

In this subsection we will describe step by step a feature extraction method called Lip Geometry Estimation (LGE). Using image filtering techniques and based on a statistical

interpretation of the results from the filters it directly estimates the geometry of the mouth. However, this technique is unique because it does not rely on any a-priori geometrical lip model.

As the first step of the processing pipeline we have to locate the face and then the mouth of the speaker. The detection of the Region of Interest (ROI) removes unnecessary areas from the image which is very important from at least two reasons: firstly the processing time is greatly reduced and secondly many possible unwanted artifacts can be avoided. For this we use the Viola-Jones algorithm for object detection. This classifier uses a new method for detecting the most representative Haar like features using a learning algorithm based on AdaBoost. It combines the weak classifiers using a "cascade" approach which corroborated with a fast method for computing the Haar-like features allows for high speed and very low false-negative rates. The next step in the process is to somehow detect which pixels belong to the lips.

Fortunately, now, because the input image contains only the mouth area and since the lips have a distinct coloring we can extract the lip's pixels without the need for complicated object recognition techniques. In order to utilize this fact, we need to apply some sort of lip-selective filter to the image. In our current research we use several different filters depending on the illumination conditions and the quality of the recorded video sequences.After extracting the appropriate ROI, the filtered image is treated as a bivariate distribution I(X, Y) (after normalization). The mean of this distribution: [EX, EY] approximates accurately and in a stable way the center of the mouth. Using this value, we transform the image into polar coordinates. Next we take the mean and variance values for any angle α. As the image is discrete rather that continuous, all of the values are obtained from summation rather than integration, so we only operate on estimations of those values, namely M(α) and $\sigma^2(\alpha)$. The vectors resulting from sampling of those functions for one of the video sequences can be seen in Figure 2.1.

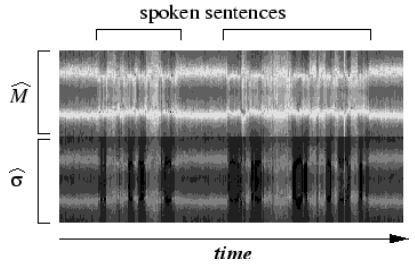

Figure 2.1: Pairs of M(α) and σ2(α) vectors extracted from a video sequence. The periods of silence around two spoken sequences can be seen clearly.

2.1.2 Mouth motion estimation based on optical flow analysis

Until now in the domain of lip reading and audio-visual speech recognition the optical flow analysis was used as raw data, or as a method to measure the global movement of the speaker face. The variances of the horizontal and vertical components of flow vectors were used as visual features for silence detection, in the cases when the noise in the audio modality was not allowing for an accurate decision.

Even though the results were very promising we argue that using the optical flow only as a global measure much of the information about speech is discarded. We propose here a method that based on the optical flow better describes the actual speech. Our method measures the lip movement on the contour of the mouth. The first step of the method tries to accurately detect the center of the speaker mouth. Since the LGE method provides a good way for detecting the center of the mouth we used that approach again for the present method. Also the detection of the appropriate region of interest is highly appreciated. Since the optical flow vectors can be computed in every region of the image we need to restrict the searching space in order to exclude unnecessary regions.

2.1.3 Intensity based features

The shape of the lips is not the only determinant of a spoken utterance. There are some other important factors such as the position of the tongue, teeth etc. Some of them can be observed in the video sequence, the others not. It is essential in the case of lip reading to extract from the visual channel as much information as possible about the utterance being spoken. We propose therefore to augment the visual features extracted until now with a few simple intensity related features.

It would probably be possible to track the relative positions of the teeth and tongue with respect to the lips. The tracking accuracy would be limited by the fact that the visibility of lips and tongue is normally very poor. Such a task would also be too complex and therefore infeasible for a lip reading application.

There are however some easily traceable features that can be measured in the image which relate to the positions and movements of the crucial parts of the mouth. The teeth for example are much brighter than the rest of the face and can therefore be located using a simple filtering of the image intensity.

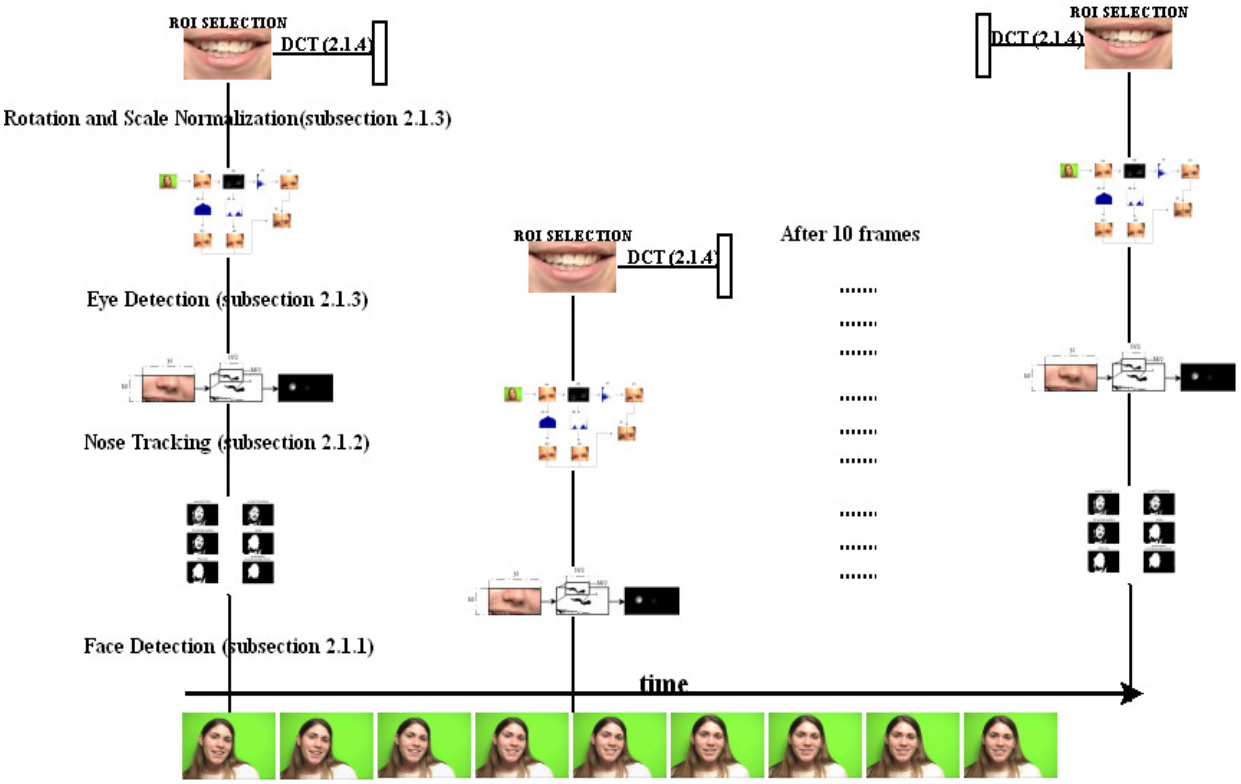

Figure 2.2: Visual Feature Extraction Schematic diagram

2.2 Visual Front-End

This section describes how visual features are extracted from each video frame and get processed to form a set of vectors, which exploit visual information and enhance audio features. An efficient visual front end system which is able to track and locate the speaker's face and mouth region-of-interest (ROI) was developed using the Viola-Jones algorithm. Initially, the classifiers for the face, eyes and mouth were developed using the OpenCV libraries. The overall visual frond-end system was developed using Microsoft Visual C++ to detect the face and extract the mouth ROI.

Given the video of a speaker, initially the system detects the face using the face classifier. Once the face was located, we then locate the eyes and based on these locations, the face was similarity normalized (i.e. normalized with respect to scale, rotation and translation) based on an inter-ocular distance of 32 pixels. We then applied a mouth classifier and from that we extracted a 32_32 ROI. Overall accuracy for the ROI detector was 92.4% within a separate validation set where 11 facial feature points (i.e. eyes, nose, lip corners, chin etc.) were manually labeled.

The following steps describe the process of the visual feature extraction and figure (2.2) depicts its schematic:

1. Face Detection (every 10 frames) and nose template re-formation (every 10 frames). Chroma based face detection is used for the face detection.
2. Template matching to track the nose (in every frame). Here nose tracker for detecting nose positions, and being able to extract the ROI using the nose tip as reference.
3. Lip region estimation.
4. Eye detection, estimation of the head inclination, scaling estimation. Eye localization and template matching is applied for the accurate detection of the iris centers.
5. Image rotation and scaling correction and ROI determination.
6. DCT feature extraction.

2.3. Visual feature extraction

Following the ROI extraction, a feature mean normalization step was used to remove any redundant information, such as illumination or speaker variances by subtracting the mean image over the utterance. This approach is very similar to cepstral mean subtraction (CMS) in the audio domain. A two-dimensional separable discrete cosine transform (DCT) was then applied to the mean removed image resultant image. The top 100 features energy components were selected to capture the static information. Subsequently, in order to incorporate dynamic speech information, seven of these neighboring static feature vectors over 3 adjacent frames were concatenated, and were projected via an inter-frame linear discriminant analysis (LDA) step to yield a 40-dimensional "dynamic" visual feature vector, extracted at the video frame rate of 30 Hz. The classes used for LDA matrix calculation were the HMM states, based on forced alignment using an audio-only HMM. These features were then used to train up a 9 state left-to-right HMM word models using 8 mixtures.

3. Audio-Visual Integration for Speech Recognition

3.1. Integration Methods

One of the main challenges in AV-ASR systems is the audio-visual information integration problem. The main issues in information integration are, (a) the class conditional dependence assumption made across streams, (b) the level (e.g. frame, phone, word) of integration, and (c) the kind (e.g. feature, partial likelihood, partial decision) of integration.

3.1.1 Early Integration

Early Integration (EI) or Feature Fusion assumes class-conditional dependence between streams and frame synchronous information integration. Audio and visual features are computed from the acoustic and visual speech respectively and they are combined before the recognition experiment. Since the two set of features correspond to different feature spaces, they may differ in their characteristics.

3.1.2 Late Integration

Late Integration (LI) or Decision Fusion incorporates separate recognizers for audio and video channels and then combines the outputs of the two recognizers to get the final result. The final step of combining the two outputs is the most important step in this approach, as it has to deal with the issues of orthogonality between the two channels and the reliability of the channels.

3.2 Stream Reliability Estimation

3.2.1 Entropy of Posteriori Probabilities

This approach estimates stream confidence by computing the mean entropy over all the class-conditional a posteriori probabilities that occur in stream, for a specific frame in time. The computed entropy values are inversely proportional to the stream confidence. When all classes have almost the same probabilities then, entropy reaches its max value and subsequently the stream is considered to be totally unreliable.

3.2.2 N-best Log-Likelihood Difference/Dispersion

The distribution of the N-best log likelihoods in time t, consists a measure for the class discrimination in stream s, based on the observation. A reasonable choice for capturing such discrimination is either to compute the log likelihood difference average between the biggest log likelihood and the rest N-1 best values (see Eq. 3.1) or alternatively to find the differences between each possible likelihood couple of the N-best and then take an average (see Eq. 3.2). Suppose that $R_{s;t;n} = \log P(o_{s;t} \mid c_{s;t;n})$ is the n^{th} best class-conditional log likelihood in stream s, in time frame t. Subsequently, $R_{s;t;1} = \log P(o_{s;t} \mid c_{s;t;1})$ corresponds to the best log likelihood.

$$Diff_{s,t} = \frac{1}{N-1}\sum_{n=2}^{N}(R_{s,t,1} - R_{s,t,n}) \tag{3.1}$$

$$Disp_{s,t} = \frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{n'=n+1}^{N}(R_{s,t,n} - R_{s,t,n'}) \tag{3.2}$$

3.2.3 K-means Clustering Method

In the presence of modeling or estimation error in a multi-stream classification problem, stream weights can decrease total classification error. For a two-stream, two-class problem, the optimal weights have been shown to be inversely proportional to the single stream pdf estimation error.

$$\frac{s_1}{s_2} = \frac{\sigma_{s2}^2}{\sigma_{s1}^2}, \tag{3.3}$$

where σ^2_s defines the variance of the pdf estimation error for stream s. Subsequently, the optimal stream weights are inversely proportional to the variance of the pdf estimation error; here we show the feature extraction of the visual extraction in the video live streaming.

In the same work, it has been also proved that assuming equal pdf estimation error variances for the two streams, the next relation stands,

$$\frac{s_1}{s_2} \approx \frac{p(x_2 \mid c_1)}{p(x_1 \mid c_1)} \; for \; 0.5 \leq \frac{p(x_2 \mid c_1)}{p(x_1 \mid c_1)} \leq 1.5, \tag{3.4}$$

3.3 Weighting Schemes for Audio-Visual Fusion

In the previous, it was described in what ways it is possible to fuse audio with visual stream using HMM. We chose a K-means based, unsupervised method to estimate stream weights for the classification and recognition task It is widely accepted that employing weighted product rule for combining partial decisions made from separate audio and visual recognizers, is a good choice.

Multi-stream HMM allows different levels of asynchrony between audio and visual streams. Phone-level or word-level boundaries are generally considered for fusing partial decisions.

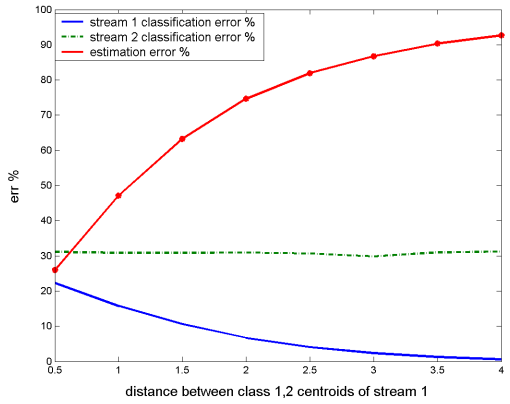

Fig 3.1 Simulation results using artificial data

3.3.1. Stream Reliability to Stream Weight Mapping

A Non-linear function f (.) is used in order to map estimated stream classification error ratio into stream weight ratio. A constant value a is also used for taking into account the class-conditional pdf estimation error difference between the two streams, as described in previous section. Note that constant a is included in the mapping function as a parameter. Overall, four types of mapping are tested in order to find which one gives better results for different types of noise and SNR levels. To compute the parameters a , b or only a for each type of mapping, we perform a fitting between the estimated and optimal weight ratios for the training data using those parametric mapping functions. After training each type of function, we test their performance over the testing data by comparing the recognition results. Weights are estimated for every sentence of ten digits but only a mean value is selected and applied to the synchronous multi-stream HMMs, in order to compare the results for different mapping functions.

4. Noise Filtering

Blocks classified as steady are filtered in time domain in order to suppress noise. In perfect case, transmittance of filter should correspond to shape of PSD (power spectral density) of the additive noise, but unfortunately exact noise model of the noise is not known a priori. Since our primary goal is not to discard noise entirely, but only assure that it won't mislead motion compensation module, we decided to find best filter characteristic empirically.

Although it would be optimal to calculate the motion estimation thresholds for each sequence separately, it would be necessary to assume that algorithm has some primal knowledge about sequence quality. Since it is difficult to distinguish noise from subtle movement and nearly impossible to do so in case of chaotic movement (e.g. waves on water), authors assume that presented de-noising algorithm requires feedback from user to avoid the unwanted video pixels that damage the quality of the video lectures in live. Filtering is the most time-consuming operation in our algorithm and we decided to incorporate IIR filter in our design. To minimize non-linearity of phase that would otherwise introduce distortions we chose low-order Butterworth type filters. Some examples are presented below in fig. 5.1-5.3.

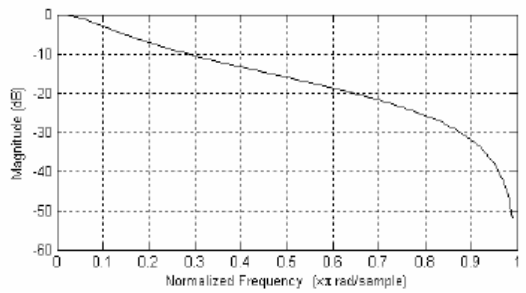

Fig 5.1 First order Butterworth filter w = 0.1.

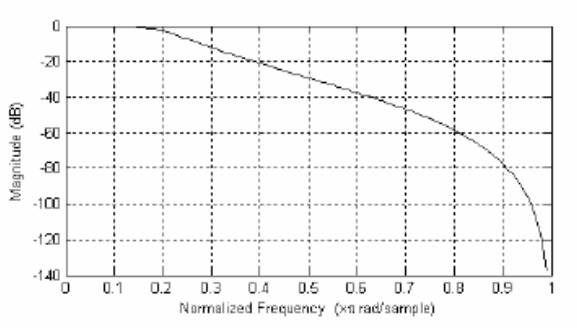

Fig 5.2 Second order Butterworth filter w = 0.2

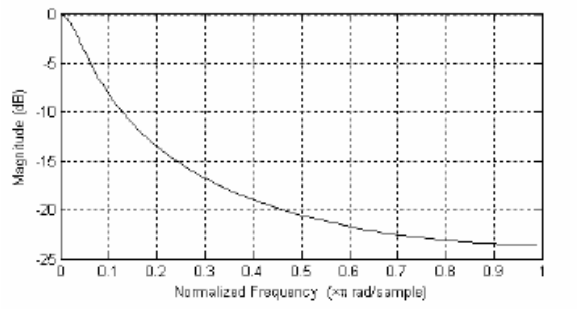

Fig 5.3 First order Butterworth filter w = 0.1, with zeroed higher nominator coefficients (except of one). Chosen for further evaluation as an effect of excellent subjective performance

5. Conclusion

The research in the domain of audio-visual speech recognition is still going on. In contrast with audio feature extraction technology, in the case of the visual modality it is still not very clear what features are more appropriate for robust lip-reading and audio-visual speech recognition. It has been show that the noise in the video lectures is reduced and the word error rate is improved. To compare the audio-visual implementation with an audio only recognizer, a multi-stream HMM recognizer was trained using "clean" data and then tested in various noisy environments which was simulated by degrading the audio with different types of noise. Definitely, the audio-visual recognizer outperformed significally both the audio-only and the visual-only recognizers.

References

[1] "Recognising Audio-Visual Speech in Vehicles using the AVICAR Database" Rajitha Navarathna1, David Dean1, Patrick Lucey1;2, Sridha Sridharan1, Clinton Fookes,2010

[2] Alin G. Chit¸u · Leon J.M. Rothkrantz · Jacek C. Wojdel · Pascal Wiggers "Comparison Between Different Feature Extraction Techniques for Audio-Visual Speech Recognition".

[3] "Temporal noise reduction for preprocessing of video streams in monitoring systems". Olgierd Stankiewicz, Adam Łuczak, Antoni Roszak

[4] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: An audiovisual speech corpus in a car environment," In Proc. Interspeech 2004, pp. 2489– 2492, Jeju Island, Korea.

[5] T. A. Ranney, W. Garrott, and M. Goodman, "NHTSA driver distraction research: past, present and future," 17th International Technical Conference on the Enhanced Safety of Vehicles, Amsterdam, June, 2001.

[6] L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Comparison between different feature extraction techniques in lipreading applications", in Specom'2006, SpIIRAS Petersburg, 2006.

[7] J. C. Wojdel and L. J. M. Rothkrantz, "Visually based speech onset/offset detection", in Proceedings of 5th Annual Scientific Conference on Web Technology, New Media, Communications and Telematics Theory, Methods, Tools and Application (Euromedia 2000), (Antwerp, Belgium), pp. 156–160, 2000.

[8] L. J. M. Rothkrantz, J. C. Wojdel, and P. Wiggers, "Fusing Data Streams in Continuous Audio-Visual Speech Recognition", in Text, Speech and Dialogue: 8th International Conference, TSD 2005, vol. 3658, (Karlovy Vary, Czech Republic), pp. 33–44, Springer Berlin / Heidelberg, September 2005.

[9] Lucey, S., Sridharan, S., Chandran, V., September 2001. An investigation of hmm classifier combination strategies for improved audio-visual speech recognition. In: [Dalsgaard et al.(2001)Dalsgaard, Lindberg, and Benner], pp. 1185-1188.

[10] Wojdel, J. C., Rothkrantz, L. J. M., September 2001. Using aerial and geometric features in automatic lipreading. In: [Dalsgaard et al.(2001)Dalsgaard, Lindberg, and Benner], pp. 2463-2466

[11] Wiggers, P., Wojdel, J. C., Rothkrantz, L. J. M., September 2002. Medium vocabulary continuous audiovisual speech recognition. In: [Hansen and Pellom(2002)], pp. 1921-1924.

[12] Chen, T., January 2001. Audiovisual speech processing. IEEE Signal Processing Magazine, 9-21.

[13] Lucey, S., Sridharan, S., Chandran, V., December 2000. An improvement of automatic speech reading using an intensity to contour stochastic transformation. In: Barlow, M. (Ed.), Proceedings of the 8th Australian Conference on Speech Science and Technology. Australian Speech Science and Technology Association, Canberra, pp. 98-103.

[14] Visser, M., Poel, M., Nijholt, A., 1999. Classifying visemes for automatic lipreading. In: Jelinek, T., Nöth, E. (Eds.), Proceedings of TSD'99. Springer Verlag, Berlin Heidelberg, pp. 349-352.

E.S.Selvakumar received the B.Tech degree in 2011 in Information Technology from the anna university and pursuing the M.Tech degree in Information Technology from periyar maniammai university. His research of interest includes the natural language processing, data mining and parallel programming.

S.Shanmuga Priya received the B.E degree in 2004 in Computer Science and Engineering from Bharathidasan University and M.Tech degree in 2007 from Sastra University. His research of interest includes the natural language processing, data mining and parallel programming.