Two-Level Bimodal Association for Audio-Visual Speech Recognition

Jong-Seok Lee[1] and Touradj Ebrahimi[1]

[1] Multimedia Signal Processing Group
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
{jong-seok.lee, touradj.ebrahimi}@epfl.ch
http://mmspg.epfl.ch

Abstract. This paper proposes a new method for bimodal information fusion in audio-visual speech recognition, where cross-modal association is considered in two levels. First, the acoustic and the visual data streams are combined at the feature level by using the canonical correlation analysis, which deals with the problems of audio-visual synchronization and utilizing the cross-modal correlation. Second, information streams are integrated at the decision level for adaptive fusion of the streams according to the noise condition of the given speech datum. Experimental results demonstrate that the proposed method is effective for producing noise-robust recognition performance without a priori knowledge about the noise conditions of the speech data.

1 Introduction

In the field of speech-based human-computer interaction, it becomes important to utilize the acoustic and the visual cues of speech simultaneously for effective recognition of spoken language by computers. Audio-visual speech recognition (AVSR) systems which additionally observe lip movements along with acoustic speech have been proposed and shown to produce enhanced noise-robust performance due to the complementary nature of the two modalities [1]. The speakers' lip movements contain significant cues about spoken language and, besides, they are not affected by acoustic noise. Therefore, the visual speech signal is a powerful information source for compensating for performance degradation of acoustic-only recognition systems in noisy environments.

How to integrate the two modalities is an important issue in constructing AVSR systems showing good recognition performance. Generally, approaches for this can be classified into two broad categories: The first one is early integration (EI), or feature fusion, in which the features from the two signals are concatenated to form a composite feature vector and then inputted to a recognizer [2]. The other one is late integration (LI), or decision fusion, where independent recognition results for the two feature streams are combined at the final decision stage [3,4]. Each approach has its own advantages against the other one. For example, constructing an AVSR system based on the EI approach is relatively simple, while, in the LI approach, it is easy to

implement an adaptive weighting of the two modalities according to the noise condition of the given speech data for noise-robust recognition performance [5].

Since the acoustic speech signal and the visual observation of the lip movements are two complementary aspects of the speech production process, there apparently exist strong cross-modal correlation which can be extracted from temporally aligned streams of the two signals and used for noise-robust recognition. On the other hand, it is known that temporal asynchrony is involved in the audio-visual correlation structure. Unfortunately, either of the two integration approaches does not model well such characteristics of audio-visual speech: The EI approach temporally correlates the two feature streams but assumes perfect synchrony between them. In the LI approach, conditional independence of one stream upon the other one is assumed and their temporal correlation is largely ignored, so that the complementary nature of the two modalities is considered only at the decision level.

In this paper, we propose a new integration method which explicitly exploits the cross-modal correlation of audio-visual speech and, thereby, enhances performance of AVSR. Our method associates the acoustic and the visual information in two levels: First, by using the canonical correlation analysis (CCA), each feature vector is projected to a new space where the correlation between the projected features is maximized, and the resultant features are concatenated for feature-level integration. In order to consider the asynchronous characteristics of the two signals, the correlation analysis is conducted with features of multiple frames. Second, decision-level association is performed between the streams of the acoustic data and the data integrated at the feature level, which adaptively combines the streams for robustness of recognition over various noise conditions. Experimental results demonstrate that the proposed method significantly enhances noise-robustness in comparison to conventional EI and LI techniques in diverse noise conditions.

The rest of the paper is organized as follows: The following section overviews existing researches related to our work. Section 3 describes the proposed system using two-level association. In Section 4, the performance of the proposed method is demonstrated via experiments on isolated-word recognition tasks. Finally, concluding remarks are given in Section 5.

2 Related Work

2.1 Audio-Visual Information Fusion

The primary goal of AVSR, i.e. robust recognition over various noise conditions, is achieved only by an appropriate information fusion scheme. Successful audio-visual information fusion should take advantage of the complementary nature of the two modalities to produce a synergetic performance gain. On the other hand, the integrated recognition performance may be even worse than the performance of any modality if the integration is not performed properly, which is called "attenuating fusion" [1].

As briefly explained in the introduction, EI and LI are the two main categories for audio-visual integration. In the former approach, the feature vectors of the two signals are concatenated and fed into a recognizer. This scheme has an advantage of simplicity. In the latter approach, the features of each modality are independently processed by the corresponding recognizer and then the outputs of the two recognizers are integrated for the final decision. Although which approach is more suitable is still arguable, there are advantages of using LI over EI for implementing noise-robust AVSR systems. First, it is easy to adaptively control relative contributions of the two modalities to the final decision with the LI approach because the modalities are processed independently. Such an adaptive control scheme is effective for producing noise-robust recognition performance over various noise conditions [3,4,5]. Second, while the EI approach assumes a perfect synchrony of the modalities, the LI approach allows flexibility for modeling the temporal relation of the acoustic and the visual signals. Previous studies suggest that audio-visual speech is not perfectly synchronized but there exist asynchronous characteristics between them: For some pronunciations, the lips and the tongue start to move up to several hundred milliseconds before the actual speech sound is produced [6]. It was demonstrated that the lips move to their initial position about 200 ms before the onset of the acoustic speech signal [7]. Also, audio-visual speech does not require precise synchrony and there exists an intersensory synchrony window during which the performance of human speech perception is not degraded for desynchronized audio-visual signals [8].

A drawback of LI is that the audio-visual correlation is not fully utilized for recognition due to separate processing of the two signals. In this paper, we solve this problem by feature-level association of the acoustic and the visual features. Previous researches for utilizing the audio-visual correlation are reviewed in the following subsection.

2.2 Audio-Visual Correlation Analysis

Previous researches on analysis of the audio-visual correlation have done mostly for speaker recognition and speaker detection. Fisher and Darrell [9] proposed a speaker association method used in multi-speaker conversational dialog systems, where an information theoretic measure of cross-modal correspondence is derived based on a probabilistic multimodal generation model so that the highly correlated part in a video sequence with the speech signal can be detected. In [10], an audio-visual feature combination method was introduced to improve speaker recognition performance; the two feature vectors are transformed by using CCA to utilize the audio-visual correlation. The idea of transforming the feature vectors of the acoustic and the visual modalities can be also found in other work: Bredin and Chollet [11] used the co-inertia analysis to measure audio-visual correspondence and showed that their method can be used to detect replay attacks in bimodal identity verification. Slaney and Covell [12] proposed a linear projection method to measure the degree of synchronization between the acoustic and the image data based on CCA.

Researchers found that it is necessary to consider asynchronous characteristics of the acoustic and the visual modalities when the cross-modal correlation of the two

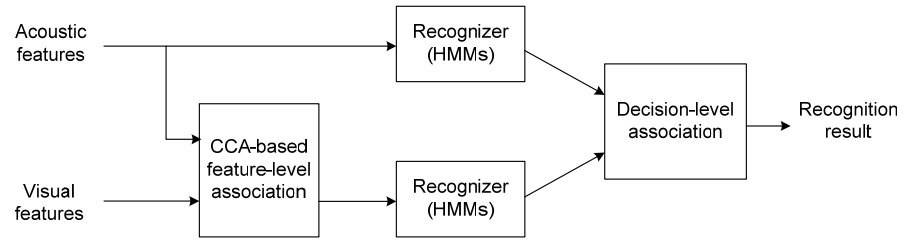

Fig. 1. Proposed system architecture

modalities is analyzed. Bregler and Konig [2] showed that the mutual information between them reaches the maximum when the visual signal is delayed up to 120 milliseconds. Based on this result, Eveno and Basacier [13] allowed a delay between the two signals up to 80 milliseconds in defining a liveness score for audio-visual biometrics. Sargin *et al*. [10] experimentally showed that feature combination with asynchrony of 40 milliseconds produces the largest cross-modal correlation and the best speaker recognition performance.

3 Proposed System

3.1 Overall System

The overall structure of our AVSR system is shown in Fig. 1. First, compact features are extracted from each signal. Then, the feature-level association takes place to combine the two feature streams with considering correlations and synchronization of the two modalities. The combined and the acoustic features are separately fed into the corresponding recognizers which are composed of hidden Markov models (HMMs). We observed that, since the feature-level association does not consider relative importance (or reliability) of the acoustic and the visual modalities, the integrated features produce worse performance than the audio-only ones for low-noise conditions while the recognition performance by using them is better than that by the visual-only ones for high-noise conditions. Therefore, the acoustic features are utilized again at the decision step in order to improve recognition performance especially for low-noise conditions. The decision-level association adaptively fuses the outputs of the two recognizers according to the noise condition of the speech signal, which produces the final recognition result.

3.2 Acoustic Feature Extraction

From the acoustic speech, we extract the popular Mel-frequency cepstral coefficients (MFCCs) [14]. While a window function having the length of 25 ms proceeds by 10 ms at a time, the 12-th order MFCCs and their temporal derivatives (delta terms) are

extracted. To reduce channel distortions contained in the speech, the cepstral mean subtraction method is applied [14].

3.3 Visual Feature Extraction

The images in the database used in this paper contain the face region around the speakers' lips. From the grayscale images, the visual features are obtained as follows [15]: First, variations within and across images such as illuminations and skin color of the speakers are reduced by left-to-right brightness balancing and pixel value normalization. Then, the two mouth corners are detected by applying thresholding. The lip region is cropped based on the found corners, and then normalized so that rotation- and scale-invariant lip region images are obtained. Next, for each pixel point, the mean pixel value over an utterance is subtracted to reduce unwanted variations across image sequences. Finally, 12-dimensional features are obtained by applying the principal component analysis (PCA) to the mean-subtracted images. Also, their delta terms are computed and used together as in the acoustic features.

3.4 Feature-Level Association

The first step for feature-level association is to make the two feature vector sequences have the same frame rate. The acoustic feature sequence usually has a higher rate than the visual one. Thus, we perform interpolation of the visual features by using cubic splines to obtain the acoustic and the visual features of the same frame rate (100 Hz in our case).

Let $\mathbf{x}_t$ and $\mathbf{y}_t$ be the N_A-dimensional acoustic feature vector and the N_V-dimensional visual one after interpolation at time t, respectively. In order to consider the correlation between them, CCA is performed with $\mathbf{x}_t$ and $\mathbf{y}_{t-\tau:t+\tau}=[\mathbf{y}_{t-\tau}, \mathbf{y}_{t-\tau+1}, \ldots, \mathbf{y}_{t+\tau}]$ which is the collection of the visual feature vectors within a window having the length of $2\tau+1$. The parameter τ determines the length of the window. In this paper, we use a symmetric window for simplicity. Correlation analysis for multiple frames has the following implications: First, it can simultaneously deal with various degrees of audio-visual asynchrony which may be different for different pronunciations. Second, the correlation between neighboring frames is considered and thus dynamic characteristics of speech can be captured. It is known that such inter-frame correlation is important for noise-robust human speech understanding [16].

The objective of CCA is to find the two transformation matrices H_A and H_V for $\mathbf{x}_t$ and $\mathbf{y}_{t-\tau:t+\tau}$, respectively, which maximize the correlation of the transformed features $\mathbf{u}_t$ and $\mathbf{v}_t$:

$$\begin{aligned} \mathbf{u}_t &= H_A^T \mathbf{x}_t \\ \mathbf{v}_t &= H_V^T \mathbf{y}_{t-\tau:t+\tau} \end{aligned} \quad . \tag{1}$$

Specifically, the first columns of H_A and H_V are obtained by solving the following maximization problem:

$$\mathbf{h}_{A1}, \mathbf{h}_{V1} = \arg\max_{\mathbf{h}_A, \mathbf{h}_V} \frac{E\left[(\mathbf{h}_A^T \mathbf{x}_t)(\mathbf{h}_V^T \mathbf{y}_{t-\tau:t+\tau})\right]}{\sqrt{E\left[(\mathbf{h}_A^T \mathbf{x}_t)^2\right] E\left[(\mathbf{h}_V^T \mathbf{y}_{t-\tau:t+\tau})^2\right]}} \quad . \tag{2}$$

The solution of the above problem, which can be found by solving an eigenvalue problem, forms the first pair of canonical basis vectors. Then, the second pair is obtained for the residuals in the same manner after the components along the first basis vectors are removed from the original data. This procedure is iteratively applied so that the extracted basis vectors compose the transformation matrices H_A and H_V. The dimension of $\mathbf{u}_t$ and $\mathbf{v}_t$, N, is given by the minimum between the dimensions of $\mathbf{x}_t$ and $\mathbf{y}_{t-\tau:t+\tau}$, i.e. N=min$\{N_A, (2\tau+1)N_V\}$. The columns of H_A and H_V, $\{\mathbf{h}_{Ai}\}$ and $\{\mathbf{h}_{Vi}\}$, i=1,2,…,N, form a set of orthonormal basis vectors for each data. More details of a general description of CCA can be found in [17].

The two transformed feature vectors, $\mathbf{u}_t$ and $\mathbf{v}_t$, are concatenated for feature-level association. Since continuous HMMs having Gaussian mixture models with diagonal covariance matrices are used for recognition, the concatenated features are transformed further by using the maximum likelihood linear transform (MLLT) [18] so that each component of the feature vector is uncorrelated with each other.

3.5 Decision-Level Association

The decision-level association is performed by a weighted sum of the outputs of the two recognizers (i.e. HMMs): For a given audio-visual datum O, the recognized utterance u* is given by [3,4]

$$u^* = \arg\max_i \left\{ \gamma \log P(O \mid \lambda_A^i) + (1-\gamma) \log P(O \mid \lambda_{AV}^i) \right\} \quad , \tag{3}$$

where λ_A^i and λ_{AV}^i are the HMMs for the acoustic and the feature-level associated features of the i-th class, respectively, and $\log P(O \mid \lambda_A^i)$ and $\log P(O \mid \lambda_{AV}^i)$ are their outputs, i.e. log-likelihoods. The weighting factor γ, whose value is between 0 and 1, determines how much each recognizer contributes to the final recognition result. It is necessary to generate an appropriate value of γ for each given audio-visual datum to produce noise-robust recognition performance. For this, we use the neural network-based method [19] where a feedforward neural network (NN) receives the "reliability" measures of the two recognizers as its inputs and produces an appropriate weighting factor as its output.

The reliability of each data stream can be measured from the corresponding HMMs' outputs. When the acoustic speech does not contain noise, the acoustic HMMs' outputs show large differences, which implies large discriminability between the classes. On the contrary, the differences become small for data containing much noise. Among various possible definition of the reliability measure based on this observation, we use the following one which has been shown to produce the best performance [5,20]:

$$S = \frac{1}{C-1}\sum_{i=1}^{C}\left\{\max_{j}\log P(O \mid \lambda^{j}) - \log P(O \mid \lambda^{i})\right\}\ , \tag{4}$$

where C is the number of classes. In other words, the reliability of a data stream is defined by the average difference between the maximum log-likelihood and the other ones computed from the HMMs for the stream.

The NN is trained so as to model the input-output mapping between the two reliabilities and the proper integration weight so that it works as an optimal weight estimator. Training is done by the following steps: First, we calculate the reliabilities of the two recognizers' outputs for the training data of a few selected noise conditions. We use ∞ dB, 20 dB, 10 dB and 0 dB speech data corrupted by white noise. Then, for each datum, the weight value for correct recognition, which appears as an interval, is searched exhaustively. For a low-noise condition, a relatively large interval of the weight produces the correct recognition results because of large differences between the HMMs' outputs; when the speech contain much noise, the interval of the weight for correct recognition becomes small. Finally, the NN is trained with the pairs of the reliabilities and the found weight values. When a test datum of unknown noise condition (which may not be considered during training) is presented, the NN produces an estimated proper weight for the datum via its generalization capability.

4 Experiments

The performance of the proposed method is demonstrated via experiments on the isolated-word audio-visual speech database which contains sixteen Korean city names [19]. For each word, three utterances of 56 speakers (37 males and 19 females) are recorded. It should be noted that the database includes more significant amounts of data for speaker-independent recognition experiments than many of the previously reported databases [21,22,23].

The acoustic signal is recorded at the rate of 32 kHz, which is downsampled to 16 kHz for feature extraction. The 12-th order MFCCs and their delta terms are extracted, as explained in Section 3.2. The visual signal contains the face region around the speakers' mouths and is recorded at the rate of 30 Hz. Twelve features based on PCA and their delta terms are extracted for each frame (Section 3.3).

The recognition task is performed in a speaker-independent manner. To increase reliability of the experiments, we use the jackknife method where the data of 56 speakers are divided into four parts and we repeat the experiments four times with the data of three parts (42 speakers) for training and the remaining part (14 speakers) for test.

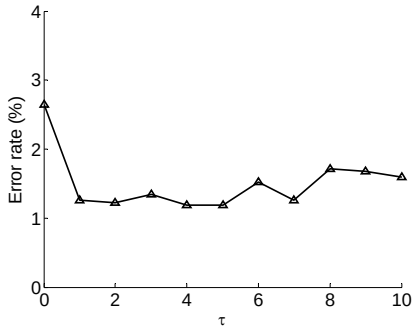

Fig. 2. Recognition performance in error rates (%) with respect to the value of τ for clean audio-visual speech data

For simulating noisy conditions, we chose four real-world noise data from the NOISEX-92 database [24]: white noise (WHT), F-16 cockpit noise (F16), factory noise (FAC) and operation room noise (OPS). Each noise signal is added to the clean acoustic speech to produce speech data of various signal-to-noise (SNR) values.

For recognizers, left-to-right continuous HMMs are used. We use the whole-word model which is a standard approach for small vocabulary recognition tasks. The number of states in each HMM is set to proportional to the number of the phonetic units of the corresponding word. We use three Gaussian functions for the Gaussian mixture model in each state.

4.1 Results of Feature-Level Association

In the correlation analysis method presented in Section 3.4, it is necessary to determine an appropriate window size τ. We determine its value by examining recognition performance for clean speech. Fig. 2 shows the recognition error rate with respect to the value of τ. In the figure, τ=0 means that CCA is performed for perfectly synchronized audio-visual features without using the temporal window. It is observed that, by using τ larger than 0, we can obtain reduced error rates compared to the case of τ=0. This implies that considering audio-visual asynchrony by using the temporal window in CCA improves recognition performance. The best performance is obtained when τ=5, which is used through our experiments henceforce.

Next, we compare various EI techniques including the proposed one. In Fig. 3, the proposed method explained in Section 3.4 (noted by "CCA(w)+MLLT") is compared with three other methods: "EI" means the conventional feature fusion method where concatenated acoustic and visual feature vectors are used for recognition. "EI+MLLT" indicates the case where MLLT is additionally applied to the "EI" features for modeling with diagonal covariance matrices in HMMs. "CCA(d)+MLLT" is a variant of the method presented in [10]; in this method, CCA is performed with desynchronized features (i.e. $\mathbf{x}_t$ and $\mathbf{y}_{t-\tau}$) and the transformed features are concatenated for recognition. For fair comparison, we additionally applied MLLT to the projected features. Here, the amount of the time delay is set to four as in [10].

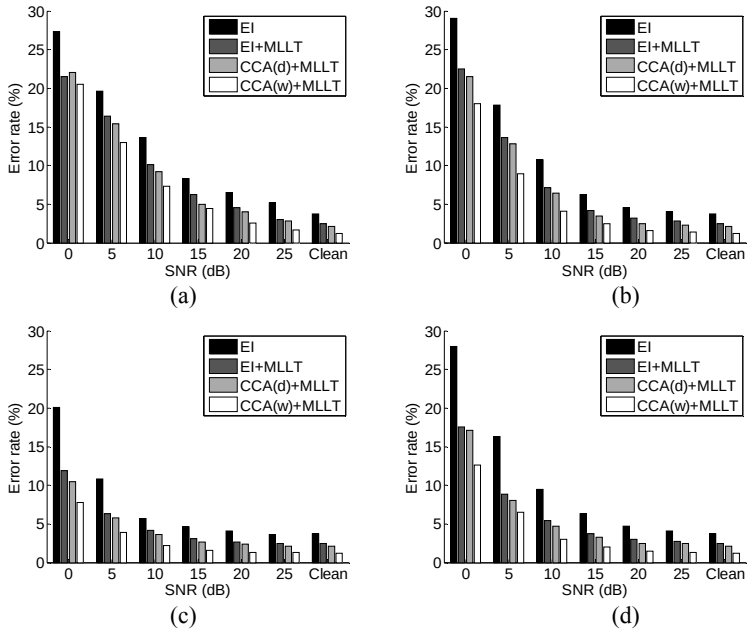

Fig. 3. Comparison of EI techniques for (a) WHT, (b) F16, (c) FAC and (d) OPS

From the figure, it is observed that the proposed method shows the best performance over various conditions. Although the recognition performance is improved by modeling asynchrony by using a relative delay between the two data sequences (i.e. "CCA(d)+MLLT") in comparison to the case without considering the delay ("EI+MLLT"), the fixed delay is not sufficient for modeling the asynchrony and thus "CCA(d)+MLLT" is outperformed by our method considering asynchrony over multiple frames.

4.2 Results of Two-Level Association

Table 1 shows the performance of the proposed AVSR system in comparison to that of unimodal and conventional bimodal recognition systems. For the case of LI, the decision-level adaptive fusion method explained in Section 3.5 is used. Note that the visual-only recognition is not influenced by acoustic noise and thus constant for all noise conditions. On the other hand, the performance of the audio-only recognition is significantly degraded when the acoustic signal contains much noise. From the results of bimodal recognition (i.e. EI, LI and the proposed method), it is observed that using bimodal information reduces the error rate significantly for low SNR conditions compared to the audio-only recognition. However, EI has a defect that its performance for high SNRs is worse than that of the audio-only recognition because EI does not control proper amounts of relative contributions of the two modalities according to the noise condition. For the same reason, EI is always outperformed by LI which adap-

Table 1. Performance of unimodal and bimodal recognition in error rates (%)

Noise	SNR (dB)	Video-only	Audio-only	EI	LI	Proposed
Clean		22.0	0.9	3.8	1.2	0.3
WHT	25		2.3	5.2	1.6	0.6
	20		5.8	6.5	2.7	1.4
	15		14.7	8.3	4.5	3.1
	10		29.5	13.6	8.5	6.5
	5		51.5	19.6	13.7	12.9
	0		74.2	27.4	19.7	21.4
F16	25		1.1	4.0	1.2	0.6
	20		1.4	4.5	1.4	0.6
	15		3.2	6.3	1.8	1.1
	10		10.9	10.8	4.5	2.4
	5		35.9	17.8	9.4	7.3
	0		81.6	29.1	19.5	18.8
FAC	25		1.0	3.7	1.2	0.4
	20		1.0	4.1	1.3	0.4
	15		1.5	4.7	1.5	0.5
	10		2.9	5.7	2.3	0.9
	5		9.3	10.8	4.0	2.1
	0		45.3	20.1	9.7	6.0
OPS	25		1.0	4.1	1.2	0.4
	20		1.6	4.8	1.3	0.6
	15		2.8	6.3	1.9	0.6
	10		9.0	9.5	3.5	1.5
	5		32.3	16.3	8.3	4.6
	0		80.8	27.9	19.2	12.6

tively adjusts the degrees of the relative contributions of the modalities. Finally, the proposed method shows the best performance among the unimodal and the bimodal recognition schemes. On average for all conditions, the relative error reduction by the proposed method over LI is 45.2%.

5 Conclusion

We have proposed a two-level association method for AVSR in which audio-visual correlation and asynchrony are considered at the feature level and the adaptive fusion of the multimodal information is performed at the decision stage. The experimental results demonstrated that the proposed correlation analysis over multiple frames can effectively exploit the audio-visual correlation present in an asynchronous manner. Moreover, it was shown that the proposed two-level association method consistently produces improved robustness over various noise conditions in comparison to the conventional unimodal and bimodal recognition schemes. An advantage of the proposed method is that it does not require a priori knowledge about the noise condition of the given audio-visual datum for robust recognition performance.

While a temporal window of a fixed length was used in our correlation analysis and its effectiveness was shown through the experiments, using a window having an utterance-dependent length may be more beneficial because an appropriate window length may vary for each utterance. Such a variable window length can be determined via statistical and linguistic analysis of audio-visual speech or optimization through training. Further study in this direction would be desirable.

Investigating the validity of the proposed method for diverse recognition tasks would be also desirable. For example, connected-word or continuous speech recognition tasks can be considered, whereas this paper addressed an isolated word recognition task. In such cases, there exist unmanageably many possible word or phoneme sequence hypotheses to be considered for the decision-level association. Solutions for this could be to consider only N-best hypotheses from each data stream and test 2N combined pairs [7], or to incorporate the adaptive weighting scheme into joint modeling of the streams by using complex models such as multi-stream HMMs [4]. Also, while we simulated the noisy conditions by adding noise to the clean speech, it would be interesting to examine with speech data recorded in noisy environments how the Lombard effect affects the performance of the proposed method, especially the correlation analysis performed in the feature-level association.

Another possibility of audio-visual association other than the feature-level and the decision-level associations is to analyze and exploit the audio-visual correlation at the signal level. This could lead to acoustic speech enhancement performed prior to feature extraction [25] and be used in conjunction with the proposed AVSR scheme.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2011) under grant agreement no. 216444 (PetaMedia), and the Swiss NCCR Interactive Multimodal Information Management (IM2).

References

1. Chibelushi, C.C., Deravi, F., Mason, J.S.D.: A Review of Speech-Based Bimodal Recognition. IEEE Trans. Multimedia 4 (2002) 23-37
2. Bregler, C., Konig, Y.: 'Eigenlips' for Robust Speech Recognition. Proc. ICASSP, Adelaide, Australia, (1994) 669-672
3. Rogozan, A., Deléglise, P.: Adaptive Fusion of Acoustic and Visual Sources for Automatic Speech Recognition. Speech Commun. 26 (1998) 149-161
4. Dupont, S., Luettin, J.: Audio-Visual Speech Modeling for Continuous Speech Recognition. IEEE Trans. Multimedia 2 (2000) 141-151
5. Lee, J.-S., Park, C.H.: Adaptive Decision Fusion for Audio-Visual Speech Recognition. In: Mihelič, F., Žibert, J. (eds.): Speech Recognition, Technologies and Applications, I-Tech, Vienna Austria (2008a) 275-296
6. Benoît, C.: The Intrinsic Bimodality of Speech Communication and the Synthesis of Talking Faces. In: Taylor, M.M., Nel, F., Bouwhuis, D. (eds.): The Structure of Multimodal Dialogue II, John Benjamins, Amsterdam, The Netherlands (2000) 485-502

7. Meyer, G.F., Mullligan, J.B., Wuerger, S.M.: Continuous Audio-Visual Digit Recognition using N-Best Decision Fusion. Information Fusion 5 (2004) 91-101
8. Conrey, B., Pisoni, D.B.: Auditory-Visual Speech Perception and Synchrony Detection for Speech and Nonspeech Signals. J. Acoust. Soc. Amer. 119 (2006) 4065-4073
9. Fisher III, J.W., Darrell, T.: Speaker Association with Signal-Level Audiovisual Fusion. IEEE Trans. Multimedia 6 (2004) 406-413
10. Sargin, M.E., Yemez, Y., Erzin, E., Tekalp, A.M.: Audiovisual Synchronization and Fusion using Canonical Correlation Analysis. IEEE Trans. Multimedia 9 (2007) 1396-1403
11. Bredin, H., Chollet, G.: Audiovisual Speech Synchrony Measure: Application to Biometrics. EURASIP J. Advances in Signal Processing 2007 (2007) Article ID 70186, 11 pages
12. Slaney, M., Covell, M.: FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. In: Leen, T.K., Dietterich, T.G., Tresp, V. (eds.): Advances in Neural Information Processing Systems, Vol. 13. MIT Press, Cambridge, Mass, USA (2001) 814-820
13. Eveno, N., Besacier, L.: Co-Inertia Analysis for "Liveness" Test in Audio-Visual Biometrics. Proc. Int. Symposium on Image and Signal Processing and Analysis, Zagreb, Croatia (2005) 257-261
14. Huang, X., Acero, A., Hon, H.-W.: Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, Upper Saddle River, NJ, USA (2001)
15. Lee, J.-S., Park, C.H.: Training Hidden Markov Models by Hybrid Simulated Annealing for Visual Speech Recognition. Proc. IEEE Int. Conf. Systems, Man, Cybernetics, Taipei, Taiwan (2006) 198-202
16. Hermansky, H.: Exploring Temporal Domain for Robustness in Speech Recognition. Proc. Int. Congress on Acoustics, Trondheim, Norway (1995) 61-64
17. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical Correlation Analysis: An Overview with Application to Learning Methods. Dept. Comput. Sci., Univ. London, UK, Tech. Rep. CSD-TR-03-02 (2003)
18. Gopinath, R.A.: Maximum Likelihood Modeling with Gaussian Distributions for Classification. Proc. ICASSP, Seattle, USA (1998) 661-664
19. Lee, J.-S., Park, C.H.: Robust Audio-Visual Speech Recognition based on Late Integration. IEEE Trans. Multimedia 10 (2008b) 767-779
20. Lewis, T.W, Powers, D.M.W.: Sensor Fusion Weighting Measures in Audio-Visual Speech Recognition. Proc. 27th Australasian Conf. Computer Science, Dunedin, New Zealand (2004) 305-314
21. Movellan, J.R.: Visual Speech Recognition with Stochastic Networks. In: Tesauro, G., Touretzky, D., Leen, T. (eds.): Advances in Neural Information Processing Systems, Vol. 7. MIT Press, Cambridge, Mass, USA (1995) 851-858
22. Chibelushi, C.C., Gandon, S., Mason, J.S.D., Deravi, F., Johnston, R.D.: Design Issues for a Digital Audio-Visual Integrated Database. Proc. IEE Colloq. Integrated Audio-Visual Processing for Recognition, Synthesis, Communication, London, UK (1996) 7/1-7/7
23. Pigeon, S., Vandendrope, L.: The M2VTS Multimodal Face Database (Release 1.00). Proc. Int. Conf. Audio- and Video-based Biometric Authentication, Crans-Montana, Switzerland (1997) 403-409
24. Varga, V., Steeneken, H.J.M.: Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. Speech Commun. 12 (1993) 247-251
25. Rivet, B., Girin, L., Jutten, C.: Mixing Audiovisual Speech Processing and Blind Source Separation for the Extraction of Speech Signals from Convolutive Mixtures. IEEE Trans. Multimedia 15 (2007) 96-108