NOISY AUDIO FEATURE ENHANCEMENT USING AUDIO-VISUAL SPEECH DATA

Roland Goecke,*[,1] *Gerasimos Potamianos*,[2] and *Chalapathy Neti* [2]

[1] Computer Sciences Laboratory, Australian National University, Canberra ACT 0200, Australia
[2] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA
E-mails: [1] Roland.Goecke@anu.edu.au; [2] {gpotam,cneti}@us.ibm.com

ABSTRACT

We investigate improving automatic speech recognition (ASR) in noisy conditions by enhancing noisy audio features using visual speech captured from the speaker's face. The enhancement is achieved by applying a linear filter to the concatenated vector of noisy audio and visual features, obtained by mean square error estimation of the clean audio features in a training stage. The performance of the enhanced audio features is evaluated on two ASR tasks: A connected digits task and speaker-independent, large-vocabulary, continuous speech recognition. In both cases and at sufficiently low signal-to-noise ratios (SNRs), ASR trained on the enhanced audio features significantly outperforms ASR trained on the noisy audio, achieving for example a 46% relative reduction in word error rate on the digits task at -3.5 dB SNR. However, the method fails to capture the full visual modality benefit to ASR, as demonstrated by its comparison to discriminant audio-visual feature fusion introduced in previous work.

1. INTRODUCTION

Although current *automatic speech recognition* (ASR) systems result to remarkably high performance for a variety of recognition tasks in clean audio conditions, their accuracy degrades rapidly with increasing levels of environment noise [1]. To handle the ASR lack of robustness to noise, the use of *visual* information obtained from the speaker's face, or mouth region, has recently been proposed [2]. Visual speech is known to contain complementary information to the acoustic signal [3], and, in addition, its quality is not affected by the audio channel noise. Furthermore, the importance of the visual modality to human speech perception is well documented [4]. It is therefore not surprising that *audio-visual* ASR has been demonstrated to outperform traditional audio-only ASR in a variety of tasks and noise conditions [2].

In addition to improving ASR, the visual modality has been investigated as a means of *enhancement* [1] of noisy audio: Girin, et. al., in [5], propose estimating clean audio features (linear prediction model coefficients, and subsequently the clean audio signal) from visual speech information, whereas in [6] and [7], they consider estimating such features from audio-visual speech, when the audio channel is corrupted by noise. Such an approach proves feasible, due to the fact that audio and visible speech are produced by the same oral-facial cavity, and are therefore *correlated*. Indeed, audio feature estimation from visual input has also been demonstrated in [8], [9], among others, and used in applications such as speech coding [10] and separation of speech signals [11].

Clearly, audio-visual ASR and audio-visual speech enhancement differ in their aims and methodologies; however, one expects that the latter would also lead to improved recognition performance over the use of a noisy audio-only based ASR system. Furthermore, enhancing the noisy audio features could enable the use of clean audio statistical models for ASR over a wide variety of noisy environments, thus avoiding noise-dependent statistical model training and storage. Therefore, it is of interest to study the effects of audio-visual speech enhancement to ASR and to compare the resulting system performance to traditional audio-visual ASR. To our knowledge, such a study has not been conducted before, and it constitutes the main topic of this paper.

In particular, we follow the approach of [6], [7] to enhance noisy audio features by means of a linear filter (transform), which is applied on the concatenated vector of noisy audio and visual features. Similarly to that work, the filter is obtained by *mean square error* (MSE) estimation of the clean audio feature training data. However, when estimating the linear transform, in addition to using the *Euclidean* distance as the estimation error metric, we consider the *Mahalanobis* distance based on a speech class labeling of the feature vectors. Since we are interested in ASR, we do not consider the problem of obtaining enhanced speech from the enhanced audio features. So, instead of using linear prediction coefficients, as in [5]–[7], we use discriminant transformations of *mel-frequency cepstral coefficients* (MFCCs) [1] as audio features.

To investigate the effects of the enhancement approach to ASR, we apply the algorithm to audio-visual data, suitable for both small- and large-vocabulary speech recognition tasks, where the data audio channel is corrupted by additive *non-stationary* noise at various signal-to-noise ratios. We report ASR results based on both the enhanced and the original noisy audio features, and we compare their performance to audio-visual ASR obtained by means of the *feature fusion* approach introduced in [12]. This fusion technique uses a two-stage discriminant transform to obtain bimodal features that have the same dimensionality as the audio ones, but, of course, do not approximate the clean audio features in the MSE sense.

The paper is structured as follows: Section 2 introduces necessary notation and provides the solution to the audio feature enhancement problem by using MSE estimation based on the Euclidean or Mahalanobis distance. Section 3 reviews the audio-visual feature fusion method to which the enhancement approach performance is compared, as well as the audio and visual features used. Section 4 describes the audio-visual databases and our ASR experiments, and finally, Section 5 summarizes the paper.

*Roland Göcke would like to thank the Audio-Visual Speech Technology Group at the IBM Thomas J. Watson Research Center for the opportunity to take part in this project.

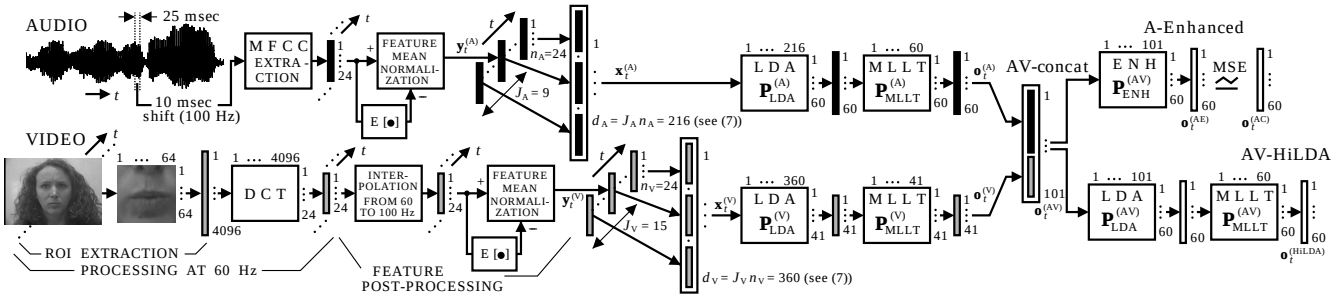

Fig. 1. Schematic diagram of the audio (*upper left*) and visual (*lower left*) front ends leading to the concatenated noisy audio-visual feature vector $\mathbf{o}_t^{(\mathrm{AV})}$, followed by noisy audio feature enhancement (*upper right*), or audio-visual discriminative feature fusion (*lower right*). See also (1), (2), and (7)-(9).

2. NOISY AUDIO FEATURE ENHANCEMENT

2.1. Problem Statement and Notation

Given an audio-visual sequence of the speaker's face, let us denote *time-synchronous* audio and visual features, extracted from it at time t, as discussed in Section 3, by [1] $\mathbf{o}_t^{(\mathrm{A})}$ and $\mathbf{o}_t^{(\mathrm{V})}$, which are vectors of dimensionality D_{A} and D_{V}, respectively. The joint audio-visual speech information is captured by the concatenation of the two single-modality feature vectors, that we denote by

$$\mathbf{o}_t^{(\mathrm{AV})} = [\, \mathbf{o}_t^{(\mathrm{A})\top},\, \mathbf{o}_t^{(\mathrm{V})\top} \,]^\top = [\, o_{t,1}^{(\mathrm{AV})},...,\, o_{t,D}^{(\mathrm{AV})} \,]^\top \in \mathsf{R}^{\,D}, \qquad (1)$$

and is of dimension $D = D_{\mathrm{A}} + D_{\mathrm{V}}$.

In addition to the speech information, the audio feature vector $\mathbf{o}_t^{(\mathrm{A})}$ captures environment noise. We hope to remove such interference and to produce *enhanced* audio features, that we denote by $\mathbf{o}_t^{(\mathrm{AE})} \in \mathsf{R}^{\,D_{\mathrm{A}}}$, using the joint audio-visual speech information captured in vector (1). The resulting enhanced audio features can then be supplied to an ASR system, hopefully yielding improved recognition over the use of the noisy ones, $\mathbf{o}_t^{(\mathrm{A})}$.

In this work, and similarly to [6], [7], we are interested in obtaining the enhanced audio $\mathbf{o}_t^{(\mathrm{AE})}$ as a linear transformation of the joint audio-visual feature vector $\mathbf{o}_t^{(\mathrm{AV})}$, namely as

$$\mathbf{o}_t^{(\mathrm{AE})} = \mathbf{P}_{\mathrm{ENH}}^{(\mathrm{AV})}\, \mathbf{o}_t^{(\mathrm{AV})}\,, \qquad (2)$$

where matrix $\mathbf{P}_{\mathrm{ENH}}^{(\mathrm{AV})} = [\, \mathbf{p}_1^{(\mathrm{AV})}, \mathbf{p}_2^{(\mathrm{AV})},..., \mathbf{p}_{D_{\mathrm{A}}}^{(\mathrm{AV})}]^\top$ is of dimension $D_{\mathrm{A}} \times D$, its rows consisting of D-dimensional vectors $\mathbf{p}_i^{(\mathrm{AV})\top}$, for $i = 1,..., D_{\mathrm{A}}$ (see also Fig.1).

To estimate matrix $\mathbf{P}_{\mathrm{ENH}}^{(\mathrm{AV})}$, we assume that in addition to (1), *clean* audio feature vectors, denoted by $\mathbf{o}_t^{(\mathrm{AC})}$, are available for a number of time instants t in a *training* set, $\mathcal{T}$.[2] We then seek to estimate the enhancement matrix in (2), such that $\mathbf{o}_t^{(\mathrm{AE})} \approx \mathbf{o}_t^{(\mathrm{AC})}$ over the training set $\mathcal{T}$, according to the two distance metrics, discussed next.

2.2. Euclidean Distance Based Estimation

A simple way to estimate matrix $\mathbf{P}_{\mathrm{ENH}}^{(\mathrm{AV})}$ is by considering the approximation $\mathbf{o}_t^{(\mathrm{AE})} \approx \mathbf{o}_t^{(\mathrm{AC})}$ in the Euclidean distance sense. Due to (2), this is equivalent to solving D_{A} MSE estimations

$$\mathbf{p}_i^{(\mathrm{AV})} = \arg\min_{\mathbf{p}} \sum_{t\in\mathcal{T}} [\, o_{t,i}^{(\mathrm{AC})} - < \mathbf{p}\,, \mathbf{o}_t^{(\mathrm{AV})} >]^{\,2}\,, \qquad (3)$$

for $i = 1,..., D_{\mathrm{A}}$, i.e., one per row of the matrix $\mathbf{P}_{\mathrm{ENH}}^{(\mathrm{AV})}$. Equations (3) result to D_{A} systems of the Yule-Walker equations [13]

$$\sum_{j=1}^{D} [\sum_{t\in\mathcal{T}} o_{t,j}^{(\mathrm{AV})}\, o_{t,k}^{(\mathrm{AV})}]\, p_{i,j}^{(\mathrm{AV})} = \sum_{t\in\mathcal{T}} o_{t,i}^{(\mathrm{AC})}\, o_{t,k}^{(\mathrm{AV})},\ k = 1,..., D\,, \qquad (4)$$

where $p_{i,j}^{(\mathrm{AV})}$ denotes the j-th element of vector $\mathbf{p}_i^{(\mathrm{AV})}$. Gauss-Jordan elimination can be used to solve (4) [13]. Note that the left hand side coefficients of all systems (4) are independent of i, and they correspond to the audio-visual feature vector covariance matrix elements (assuming zero mean observations).

2.3. Mahalanobis Distance Based Estimation

A more sophisticated way of estimating $\mathbf{P}_{\mathrm{ENH}}^{(\mathrm{AV})}$ is by weighting each term of the sum in (3) by the inverse variance of the clean audio vector element $o_{t,i}^{(\mathrm{AC})}$, denoted by $\sigma_{t,i}^2$ [13]. Thus, (3) becomes

$$\mathbf{p}_i^{(\mathrm{AV})} = \arg\min_{\mathbf{p}} \sum_{t\in\mathcal{T}} [\, \frac{o_{t,i}^{(\mathrm{AC})} - < \mathbf{p}\,, \mathbf{o}_t^{(\mathrm{AV})} >}{\sigma_{t,i}}]^{\,2}\,, \qquad (5)$$

for $i = 1,..., D_{\mathrm{A}}$. It is not difficult to see that this is equivalent to considering a Mahalanobis type distance between vectors $\mathbf{o}_t^{(\mathrm{AE})}$ and $\mathbf{o}_t^{(\mathrm{AC})}$, under the assumption of the latter having a *diagonal* covariance. Of course, estimating $\sigma_{t,i}$, for $i = 1,..., D_{\mathrm{A}}$ and $t \in \mathcal{T}$, becomes an issue. In this work, we consider clustering the training set vectors $\mathbf{o}_t^{(\mathrm{AC})}$ into a small set of *classes* $\mathcal{C}$, such as *phones*, or alternatively, context-independent, or context-dependent *hidden Markov model* (HMM) *states* [1]. Class labels at each time instant t, denoted by $c(t) \in \mathcal{C}$, can then readily be obtained by *forced alignment* of the training set utterances using a suitably trained HMM [1]. Subsequently, clean audio feature variances can be estimated for the various classes, based on the training data.

Substituting class variances in (5), it is not hard to derive the new set of D_{A} linear equation systems, the solution of which provides $\mathbf{p}_i^{(\mathrm{AV})}$, for $i = 1,..., D_{\mathrm{A}}$, namely

$$\sum_{j=1}^{D} [\sum_{t\in\mathcal{T}} \frac{o_{t,j}^{(\mathrm{AV})}\, o_{t,k}^{(\mathrm{AV})}}{\sigma_{c(t),i}^2}]\, p_{i,j}^{(\mathrm{AV})} = \sum_{t\in\mathcal{T}} \frac{o_{t,i}^{(\mathrm{AC})}\, o_{t,k}^{(\mathrm{AV})}}{\sigma_{c(t),i}^2},\ k = 1,..., D. \qquad (6)$$

Notice, that in contrast to (4), the left hand side coefficients of systems (6) now depend on the vector element i.

[1] Throughout this work, lower-case bold letters denote *column* vectors, upper-case bold letters denote matrices, whereas $\bullet^\top$ and $<\bullet\,,\bullet>$ denote a vector transpose and inner product of two vectors, respectively.

[2] Such a scenario is plausible, for example, when the noise is additive to the audio signal, and typical noise samples are known. In such a case, clean speech training data can be artificially corrupted to obtain noisy audio features that correspond to the original clean audio features.

Recognition task	Training set			Test set		
	Utter.	Dur.	Sub.	Utter.	Dur.	Sub.
LVCSR	17111	34:55	239	1038	2:29	26
DIGITS	5490	8:01	50	529	0:46	50

Table 1. The audio-visual databases and their training and test set partitioning (number of utterances, duration (in hours), and number of subjects are depicted for each set). Two recognition tasks are considered: Continuous read speech (LVCSR) and connected digits (DIGITS).

Features	11.6 dB	3.4 dB
Noisy audio-only	2.327	8.176
Enhanced audio (Euclidean distance)	2.016	5.517
Enhanced audio (Mahal.- 22 classes)	2.260	5.761
Enhanced audio (Mahal.- 66 classes)	2.282	5.783
Enhanced audio (Mahal.-159 classes)	2.282	5.805
Audio-visual (HiLDA fusion)	1.839	3.324

Table 2. Test set WER (%) for noisy audio-only, audio-visually enhanced audio (using Euclidean or Mahalanobis distance), and audio-visual HiLDA features for the DIGITS task at two noise conditions.

3. SINGLE-MODALITY FEATURES AND AUDIO-VISUAL FEATURE FUSION

We now briefly review our basic audio-visual ASR system, introduced in [12] and depicted in Fig.1. Given the audio-visual data sequence, we first extract time-synchronous *static* audio and visual features at a rate of 100 Hz, denoted by $\mathbf{y}_t^{(s)} \in \mathsf{R}^{n_s}$, where $s = \mathrm{A}, \mathrm{V}$, respectively. The audio features are 24 MFCCs, computed over a sliding window of 25ms at a rate of 100 Hz, followed by *feature mean normalization* (FMN). The visual features are the 24 highest energy *discrete cosine transform* coefficients of a 64×64 pixel mouth *region of interest* (ROI), extracted at the video field rate (60 Hz), followed by interpolation to the audio feature rate and FMN (see Fig.1). The mouth ROI is extracted using a statistical face tracking algorithm, as discussed in [12]. Subsequently, and in order to capture dynamic speech information within each modality, we concatenate J_s consecutive static features into vectors

$$\mathbf{x}_t^{(s)} = [\, \mathbf{y}_{t-\lfloor J_s/2 \rfloor}^{(s)\top}, \ldots, \mathbf{y}_t^{(s)\top}, \ldots, \mathbf{y}_{t+\lceil J_s/2 \rceil - 1}^{(s)\top} \,]^\top, \qquad (7)$$

of dimension $d_s = J_s\, n_s$, where $s = \mathrm{A}, \mathrm{V}$. To reduce the dimensionality of the resulting vectors and improve speech class discrimination, we apply a *linear discriminant analysis* (LDA) projection on (7), followed by a rotation by means of a *maximum likelihood linear transform* (MLLT) that improves statistical data modeling. This results to *dynamic* audio and visual features

$$\mathbf{o}_t^{(s)} = \mathbf{P}_{\mathrm{MLLT}}^{(s)}\, \mathbf{P}_{\mathrm{LDA}}^{(s)}\, \mathbf{x}_t^{(s)} \in \mathsf{R}^{D_s}, \qquad (8)$$

where $s = \mathrm{A}, \mathrm{V}$, and matrices $\mathbf{P}_{\mathrm{LDA}}^{(s)}$ and $\mathbf{P}_{\mathrm{MLLT}}^{(s)}$ are of dimensions $D_s \times d_s$ and $D_s \times D_s$, respectively [12] (see also Fig. 1).

Following the audio and visual feature concatenation (1), a second stage of LDA and MLLT is applied on $\mathbf{o}_t^{(\mathrm{AV})}$ to discriminantly reduce its dimensionality. The resulting fused features

$$\mathbf{o}_t^{(\mathrm{HiLDA})} = \mathbf{P}_{\mathrm{MLLT}}^{(\mathrm{AV})}\, \mathbf{P}_{\mathrm{LDA}}^{(\mathrm{AV})}\, \mathbf{o}_t^{(\mathrm{AV})} \in \mathsf{R}^{D_{\mathrm{HiLDA}}}, \qquad (9)$$

can be fed into a traditional HMM-based ASR system. Due to the two-stage application of LDA and MLLT, the method is referred to as *hierarchical LDA* (HiLDA) and it constitutes an effective feature fusion approach for audio-visual ASR [12]. In our system, we use values $n_{\mathrm{A}} = 24$, $J_{\mathrm{A}} = 9$, $D_{\mathrm{A}} = 60$, and $n_{\mathrm{V}} = 24$, $J_{\mathrm{V}} = 15$, $D_{\mathrm{V}} = 41$, whereas $D_{\mathrm{HiLDA}} = 60$. Note that $D = 101$ (see (1)), and that $D_{\mathrm{A}} = D_{\mathrm{HiLDA}}$, i.e., the dimensionalities of the enhanced audio and HiLDA audio-visual feature vectors are *equal*.

4. AUDIO-VISUAL DATABASES AND EXPERIMENTS

Our experiments are performed on two audio-visual speech data corpora: A corpus of 50 subjects uttering *connected digit* sequences (referred to as the DIGITS recognition task), as well as, on a part of the IBM ViaVoice$^{\mathrm{TM}}$ audio-visual database [12], consisting of 265 subjects uttering continuous read speech with mostly verbalized punctuation and a 10.4 k word vocabulary, i.e., a *large vocabulary, continuous speech recognition* (LVCSR) task. In both corpora, the video contains the full frontal subject face in color, has a frame size of 704×480 pixels, is captured interlaced at a rate of 30 Hz (60 fields per second are available at half the vertical resolution), and is MPEG-2 encoded at a compression ratio of about 50:1. The audio is captured at 16 kHz in an office environment at a 19.5 dB *signal-to-noise ratio* (SNR).

The two corpora are partitioned into training and test sets, suitable for *multi-speaker* (DIGITS) or *speaker-independent* (LVCSR) recognition, as depicted in Table 1. For both tasks, non-stationary speech "babble" noise is artificially added to the audio channel at various SNR values. Subsequently, at each SNR (and task), audio enhancement matrices are computed by means of MSE estimation on the training set, using the Euclidean or the Mahalanobis distance, as discussed in Section 2. HMMs are then trained on the resulting enhanced audio, and their ASR performance is evaluated on the test set. This is benchmarked against the ASR performance of HMMs trained on audio-visual HiLDA features, as well as on the noisy audio-only front end. Note that for all three systems, the LDA and MLLT matrices (see Fig. 1) are trained on data matched to the noise condition (SNR level) under consideration. HMMs with 159 and 2808 context-dependent states are used for the DIGITS and LVCSR tasks, respectively, and a tri-gram language model is used during LVCSR decoding. Furthermore, the number of HMM Gaussian mixtures is kept approximately the same across the ASR systems trained on the enhanced audio, noisy audio, and HiLDA audio-visual front ends.

In Table 2, we report performance on the DIGITS task and two SNR conditions of the ASR systems trained on the various front ends discussed in this work. Both Euclidean and Mahalanobis distance based MSE estimation for audio enhancement is considered. In particular, in the latter case, various number of classes are evaluated, namely a 22 phone class partitioning of the training set data, as well as, a 66 context-dependent HMM state and a 159 context-independent HMM state partitioning. Notice that Euclidean distance based audio enhancement slightly outperforms the Mahalanobis based approaches, therefore, in the subsequent experiments, we only consider the Euclidean distance. Furthermore, all enhanced audio features reduce *word error rate* (WER) over noisy audio-only features, but do not reach the performance of the HiLDA audio-visual features.

The last point is reinforced in Fig.2, where the performance of Euclidean distance based enhanced audio ASR is compared to

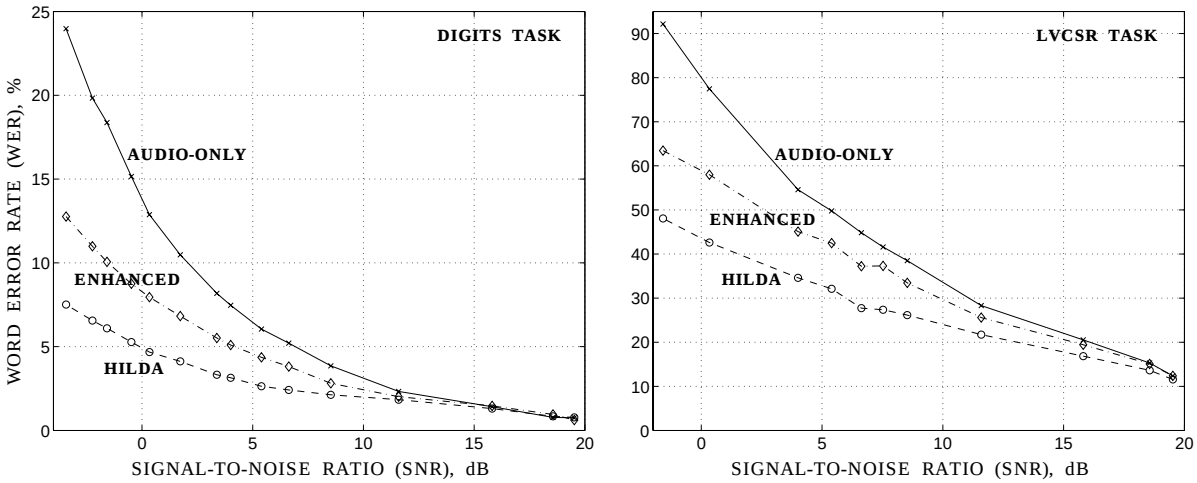

Fig. 2. Test set WER (%) for noisy audio-only, audio-visually enhanced audio, and discriminant audio-visual features, depicted against the audio channel SNR for both connected digits ASR and LVCSR tasks. WER scales differ between the two plots.

noisy audio and HiLDA based ASR, for both DIGITS and LVCSR tasks, and over a wide SNR range. At high SNRs, the WERs of the three systems do not deviate much, however at low SNRs, significant gains are obtained by the enhancement approach over noisy audio ASR. Indeed, a 46% WER relative reduction is achieved at -3.5 dB on the DIGITS task and a 31% reduction at -1.6 dB for LVCSR. However, the investigated audio enhancement technique does not capture the full benefit of the visual modality to ASR, as it is inferior to HiLDA feature fusion.

5. SUMMARY AND DISCUSSION

We investigated the effects on ASR of enhancing noisy audio features by means of audio-visual speech data. Enhancement is performed by linear filters applied on concatenated audio-visual feature vectors and obtained by MSE estimation. The method resulted to large improvements in ASR over the use of the original noisy audio features for both a small- and a large-vocabulary recognition task. Compared however to audio-visual discriminant feature fusion, the enhancement approach fared significantly worse.

We believe that the inferior performance of audio-visually enhanced speech compared to audio-visual fusion can be attributed to the following reasons: The simplicity of the linear filter used for enhancement, the non-stationarity of the noise considered in our experiments, and, likely, the very nature of the enhancement approach: By requiring that the resulting audio features approximate the clean audio ones, the visual, complementary to the audio, speech information does not get fully utilized. In contrast, feature fusion seeks an optimal, speech discriminant projection of the joint audio-visual data, without restricting the projection to be in the original audio space, thus allowing more freedom in the use of such complementary information. We expect that investigation of non-linear systems for audio-visual speech enhancement will result in further improvements to recognition performance.

6. REFERENCES

[1] J.R. Deller, Jr., J.G. Proakis, and J.H.L. Hansen, *Discrete Time Processing of Speech Signals*. Macmillan Publishing Company, Englewood Cliffs, 1993.

[2] D.G. Stork and M.E. Hennecke, eds., *Speechreading by Humans and Machines*. Springer, Berlin, 1996.

[3] D.W. Massaro and D.G. Stork, "Speech recognition and sensory integration," *American Scientist, 86*(3): 236-244, 1998.

[4] R. Campbell, B. Dodd, and D. Burnham, eds., *Hearing by Eye II*. Psychology Press, Hove, 1998.

[5] L. Girin, G. Feng, and J.-L. Schwartz, "Noisy speech enhancement with filters estimated from the speaker's lips," *Proc. Europ. Conf. Speech Comm. Techn.*, pp. 1559–1562, 1995.

[6] L. Girin, G. Feng, and J.-L. Schwartz, "Fusion of auditory and visual information for noisy speech enhancement: A preliminary study of vowel transitions," *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp. 1005–1008, 1998.

[7] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. America, 109*(6): 3007–3020, 2001.

[8] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Comm., 26*(1-2): 23–43, 1998.

[9] J.P. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models," *Proc. Work. Audio-Visual Speech Process.*, 1999.

[10] E. Foucher, L. Girin, and G. Feng, "Audiovisual speech coder: Using vector quantization to exploit the audio/video correlation," *Proc. Work. Audio-Visual Speech Process.*, 1998.

[11] L. Girin, A. Allard, and J.-L. Schwartz, "Speech signals separation: A new approach exploiting the coherence of audio and visual speech," *Proc. Work. Multimedia Signal Process.*, pp. 631–636, 2001.

[12] G. Potamianos, J. Luettin, and C. Neti, "Hierarchical discriminant features for audio-visual LVCSR," *Proc. Int. Conf. Acoust. Speech Signal Process.*, pp. 165–168, 2001.

[13] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C. The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1988.