

AUDIO-VISUAL SPEECH ENHANCEMENT WITH AVCDCN (AUDIO-VISUAL CODEBOOK DEPENDENT CEPSTRAL NORMALIZATION)

Sabine Deligne, Gerasimos Potamianos, Chalapathy Neti

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA.
emails: {deligne,gpotam,cneti}@us.ibm.com

ABSTRACT

In this paper, we introduce a **non-linear** enhancement technique called *Audio-Visual Codebook Dependent Cepstral Normalization (AVCDCN)* and we consider its use with both audio-only and audio-visual speech recognition. AVCDCN is inspired from CDCN [1] [2], an audio-only enhancement technique that approximates the non-linear effect of noise on speech with a piece-wise constant function. Our experiments show that the use of visual information in AVCDCN allows significant performance gains over CDCN.

1. AUDIO-VISUAL APPROACH TO SPEECH RECOGNITION

Although current *automatic speech recognition (ASR)* systems perform remarkably well for a variety of recognition tasks in clean audio conditions, their accuracy degrades with increasing levels of environment noise. New approaches are needed to handle the ASR lack of robustness to noise. In this paper, we propose a multi-sensor approach to ASR, where visual information, in addition to the standard audio information, is obtained from the speaker's face in a second channel. Audio-visual ASR, where both an audio channel and a visual channel are input to the recognition system, has already been demonstrated to outperform traditional audio-only ASR in noise conditions [5] [6]. In addition to audio-visual ASR, the visual modality has been investigated as a means of enhancement, where clean audio features are estimated from audio-visual speech when the audio channel is corrupted by noise [3] [4]. However, in [4] for example, the ASR performance of linear audio-visual enhancement (where clean audio features are estimated via linear filtering of the noisy audio-visual features) remains significantly inferior to the performance of audio-visual ASR. In this paper, we introduce a **non-linear** enhancement technique called *Audio-Visual Codebook Dependent Cepstral Normalization (AVCDCN)* and we consider its use with both audio-only ASR and audio-visual ASR. AVCDCN is inspired from

CDCN [1] [2], an audio-only non-linear enhancement technique which is well known in the field of ASR. In CDCN, the non-linear effect of the noise on the clean speech features is approximated with a piece-wise constant function. AVCDCN is a multi-sensor extension of CDCN that integrates the use of audio and visual features. Our experiments show that the use of visual information in AVCDCN allows significant performance gains over CDCN.

2. PRINCIPLE OF AVCDCN

Let's denote $\mathbf{x}^A(\mathbf{t})$ a cepstral vector of audio features corrupted by noise and observed at time \mathbf{t} , $\mathbf{n}(\mathbf{t})$ the unknown vector of noise features and $\mathbf{y}^A(\mathbf{t})$ the unknown vector of clean speech features that would have been observed in the absence of noise. The principle of CDCN [1] [2] is to compute an estimate $\hat{\mathbf{y}}^A(\mathbf{t})$ of $\mathbf{y}^A(\mathbf{t})$ as the expected value of $\mathbf{y}^A(\mathbf{t})$ given the observed noisy features $\mathbf{x}^A(\mathbf{t})$:

$$\hat{\mathbf{y}}^A(\mathbf{t}) = \int_{\mathbf{y}^A} \mathbf{y}^A p(\mathbf{y}^A | \mathbf{x}^A(\mathbf{t})) d\mathbf{y}^A \quad (1)$$

Using the fact that:

$$\mathbf{y}^A(\mathbf{t}) = \mathbf{x}^A(\mathbf{t}) - \mathbf{r}(\mathbf{y}^A(\mathbf{t}), \mathbf{n}(\mathbf{t}))$$

with \mathbf{r} a non linear function of both the clean audio and the noise [1], equation 1 becomes:

$$\hat{\mathbf{y}}^A(\mathbf{t}) = \mathbf{x}^A(\mathbf{t}) - \int_{\mathbf{y}^A} \mathbf{r}(\mathbf{y}^A, \mathbf{n}(\mathbf{t})) p(\mathbf{y}^A | \mathbf{x}^A(\mathbf{t})) d\mathbf{y}^A \quad (2)$$

The novelty with AVCDCN is to use the visual modality to estimate more accurately the correction term applied to $\mathbf{x}^A(\mathbf{t})$. Denoting $\mathbf{x}^{AV}(\mathbf{t})$ the vector of concatenated audio-visual features observed at time \mathbf{t} :

$$\hat{\mathbf{y}}^A(\mathbf{t}) = \mathbf{x}^A(\mathbf{t}) - \int_{\mathbf{y}^A} \mathbf{r}(\mathbf{y}^A, \mathbf{n}(\mathbf{t})) p(\mathbf{y}^A | \mathbf{x}^{AV}(\mathbf{t})) d\mathbf{y}^A \quad (3)$$

For lack of being able to compute $\mathbf{r}(\mathbf{y}^A, \mathbf{n}(\mathbf{t}))$ since the noise $\mathbf{n}(\mathbf{t})$ corrupting the speech is not known (and to avoid

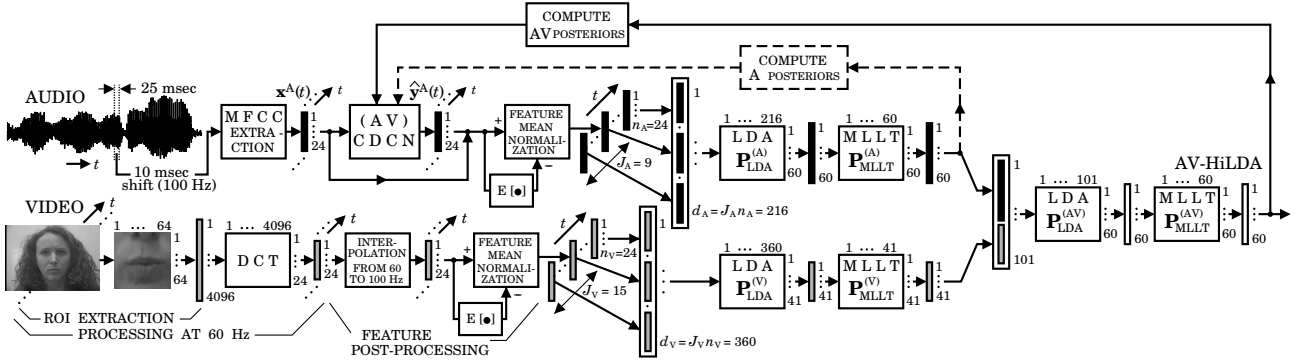


Fig. 1. Schematic diagram of the audio (*upper left*) and visual (*lower left*) front ends, followed by either audio ASR (*upper right*) or audio-visual ASR (*right*). Either AVCDCN or CDCN enhancement is applied on the audio MFCCs before feature normalization. The AVCDCN and CDCN codeword posteriors are estimated from the audio-visual and audio features respectively, after the corresponding LDA/MLLT projections.

computing an integral over \mathbf{y}^A , we approximate equation 3 with a sum computed over a pre-defined codebook of audio compensation terms $\{r_k^A\}_{k=1}^K$

$$\hat{\mathbf{y}}^A(\mathbf{t}) = \mathbf{x}^A(\mathbf{t}) - \sum_{k=1}^K r_k^A \mathbf{y}(\mathbf{k} | \mathbf{x}^{AV}(\mathbf{t})) \quad (4)$$

In our experiments, we compare AVCDCN to its audio-only counterpart, CDCN, defined by:

$$\hat{\mathbf{y}}^A(\mathbf{t}) = \mathbf{x}^A(\mathbf{t}) - \sum_{k=1}^K r_k^A \mathbf{y}(\mathbf{k} | \mathbf{x}^A(\mathbf{t})) \quad (5)$$

Note that AVCDCN and CDCN use the same set of audio compensation codewords, however AVCDCN takes advantage of the visual information to estimate the posterior distribution of the codewords. In the next section, we explain how to estimate the audio compensation codewords and their CDCN or AVCDCN posterior distributions.

3. ESTIMATION OF THE AVCDCN PARAMETERS

The posterior distribution $\{\mathbf{y}(\mathbf{k} | \mathbf{x}^{AV}(\mathbf{t}))\}_{k=1}^K$ is computed by assuming that the *probability density function* (pdf) of \mathbf{x}^{AV} is a mixture of Gaussians with priors, means and covariances $(\pi_k^{AV}, \mu_k^{AV}, \Sigma_k^{AV})_{k=1}^K$, so that by using Bayes rule:

$$\mathbf{y}(\mathbf{k} | \mathbf{x}^{AV}(\mathbf{t})) = \frac{\pi_k^{AV} \mathcal{N}(\mathbf{x}^{AV}(\mathbf{t}); \mu_k^{AV}, \Sigma_k^{AV})}{\sum_{j=1}^K \pi_j^{AV} \mathcal{N}(\mathbf{x}^{AV}(\mathbf{t}); \mu_j^{AV}, \Sigma_j^{AV})} \quad (6)$$

where \mathcal{N} refers to the Gaussian pdf. In our experiments, both the codebook of audio compensations and the pdf parameters of the noisy audio-visual features are estimated from a stereo training database consisting of clean audio

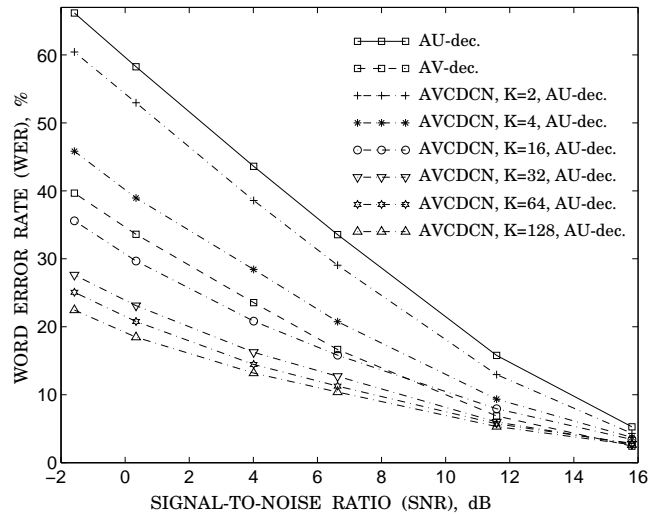


Fig. 2. Audio-only ASR using AVCDCN-enhanced features for various codebook sizes. WERs are plotted against the audio-channel SNR. For comparison, the performance of audio-only and audio-visual ASR is also depicted. All the results reported on this figure are for HMMs trained on clean data.

features $(y^A(\mathbf{t}))_{\mathbf{k}}^T$ in the first channel and of noisy audio-visual features $(x^{AV}(\mathbf{t}))_{\mathbf{k}}^T$ in the second channel. The noisy audio data in the second channel are generated by adding noise to the waveform of the clean audio features contained in the first channel. The audio compensations are computed by minimizing the expected square error between y^A and x^A over the stereo training database:

$$r_{\mathbf{k}}^A = \frac{\sum_{\mathbf{t}=1}^T (x^A(\mathbf{t}) - y^A(\mathbf{t})) p(\mathbf{k} | y^A(\mathbf{t}))}{\sum_{\mathbf{t}=1}^T p(\mathbf{k} | y^A(\mathbf{t}))} \quad (7)$$

and Maximum Likelihood (ML) estimates of the means and covariances of the noisy audio-visual features are computed as (assuming equal priors):

$$\begin{aligned} \mu_{\mathbf{k}}^{AV} &= \frac{\sum_{\mathbf{t}=1}^T x^{AV}(\mathbf{t}) p(\mathbf{k} | y^A(\mathbf{t}))}{\sum_{\mathbf{t}=1}^T p(\mathbf{k} | y^A(\mathbf{t}))} \\ \Sigma_{\mathbf{k}}^{AV} &= \frac{\sum_{\mathbf{t}=1}^T x^{AV}(\mathbf{t}) x^{AV}(\mathbf{t})^T p(\mathbf{k} | y^A(\mathbf{t}))}{\sum_{\mathbf{t}=1}^T p(\mathbf{k} | y^A(\mathbf{t}))} - (\mu_{\mathbf{k}}^{AV})^2 \end{aligned} \quad (8)$$

where T denotes transposition. The posteriors $p(\mathbf{k} | y^A(\mathbf{t}))$ are computed by assuming that the pdf of the clean audio features y^A is a mixture of Gaussians with equal priors, and with means and covariances for which ML estimates are computed with a standard expectation-maximization algorithm on the clean audio training data. In our CDCN baseline, the means and covariances of the noisy audio features are computed by replacing $x^{AV}(\mathbf{t})$ by $x^A(\mathbf{t})$ in equations 8.

4. EXPERIMENTS

4.1. Database and recognition systems

Our experiments are performed on an audio-visual corpus of 50 subjects uttering connected digit sequences. 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 30Hz (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 16kHz in an office environment at a 19.5dB *signal-to-noise ratio* (SNR). The corpus is partitioned into training (5,490 utterances, 8 hours, 50 subjects) and test (529 utterances, 0.46 hour, 50 subjects) sets for multi-speaker recognition, i.e. test speakers are also present in training. Non-stationary speech babble noise is artificially added to the audio channel at various SNR values.

Estimating the parameters for AVCDCN and CDCN requires first to compute the pdf characterizing the clean speech in the audio channel of the training set. A set of audio compensation codewords and the pdf characterizing the noisy audio-visual speech (resp. the noisy audio-only speech in the case of audio-only CDCN) are then estimated, for each

SNR condition, according to equations 7 and 8. The front end of our systems is shown on Figure 1. The audio compensation codewords $r_{\mathbf{k}}^A$ are estimated on the MFCC features before applying the audio feature mean normalization. The pdfs of the noisy speech are estimated on the features output by the audio-visual LDA/MLLT transform (AV-HiLDA, see [6]) for AVCDCN and on the features output by the audio LDA/MLLT transform for CDCN. When decoding the test set, the MFCC features are enhanced with either AVCDCN or CDCN according to eq. 4 or 5, where the posterior probabilities are computed with the features output by the LDA/MLLT transforms and the pdfs of noisy speech matching the SNR level under consideration.

The AVCDCN and CDCN enhancement strategies are evaluated for various sizes of codebooks across all SNR levels with both audio and audio-visual ASR. This is benchmarked against audio and audio-visual ASR without enhancement. Furthermore, we report on recognition experiments where the LDA/MLLT transforms producing the features sent to the decoder and the HMMs used by the decoder are either trained on the clean training data or re-trained on the enhanced noisy training data matching the SNR level under consideration. All ASR systems (trained on the clean data or re-trained on the noisy and enhanced data, with the audio and audio-visual front ends) use a set of HMMs with 159 context-dependent states and a total 3.2K Gaussians.

4.2. Results

Word Error Rates (WERs) obtained on the test set are plotted as a function of the SNR on Figures 2 and 3. Figure 2 shows how the WERs obtained with AVCDCN in an audio-only ASR scheme decreases consistently at all SNR values when the size of the codebook is increased from $K = 2$ to $K = 128$ codewords. On the same figure are also plotted the WERs obtained in an audio-only and audio-visual ASR scheme without enhancing the features. AVCDCN outperforms the audio-only ASR scheme regardless of the size of the codebook, and most interestingly, AVCDCN outperforms also the audio-visual ASR scheme for codebooks of at least 16 codewords.

Figure 3 compares AVCDCN and CDCN in both audio and audio-visual ASR schemes. To maintain the clarity of the figures, only the WERs obtained with codebooks of 128 codewords are plotted, but the conclusions drawn here apply for all sizes of codebook. Figure 3.(a) shows WERs obtained with the recognition systems trained on the original clean training data. Figure 3.(b) shows WERs obtained with the recognition systems re-trained on: (i) the noisy training data matching the SNR level under consideration when no enhancement is used, (ii) the CDCN-enhanced or AVCDCN-enhanced noisy training data matching the SNR level under consideration when either CDCN or AVCDCN is used. When systems are not re-trained (Figure 3.(a)),

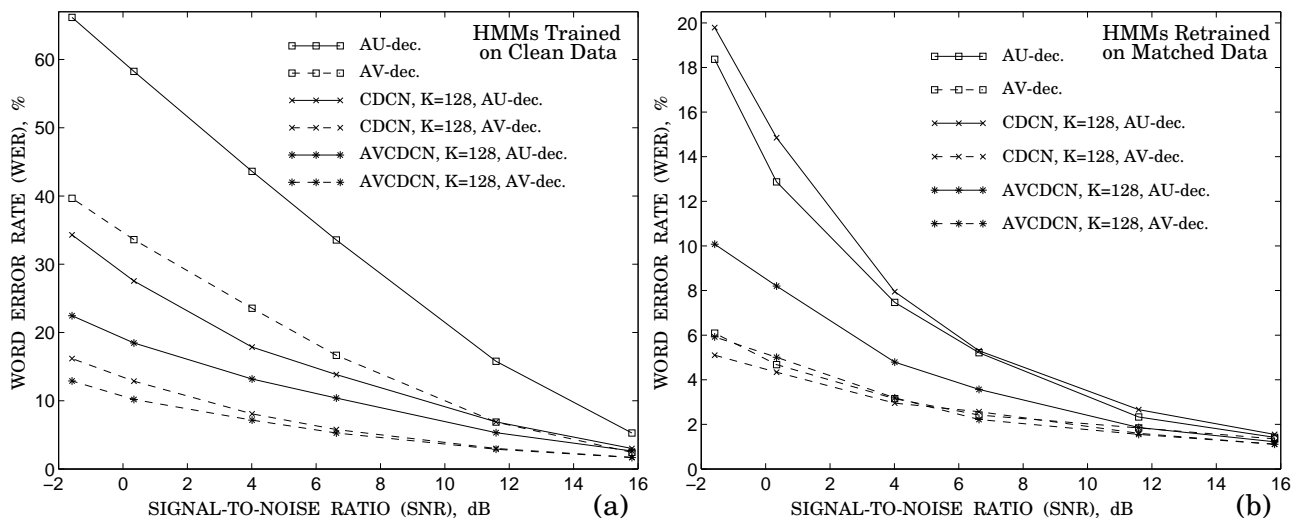


Fig. 3. Audio-only and audio-visual ASR on noisy features, CDCN or AVCDCN-enhanced features using: (a). HMMs trained on clean data, (b). HMMs trained on the noisy or enhanced features matching the SNR level under consideration.

AVCDCN performs significantly better than CDCN in both the audio and audio-visual ASR schemes. Also, the performance gains obtained with AVCDCN and with audio-visual ASR add up since AVCDCN combined with audio-visual ASR significantly outperforms both audio-visual ASR, and, AVCDCN combined with audio-only ASR. Re-training the systems (Figure 3.(b)) improves the performances of all strategies. AVCDCN still performs significantly better than CDCN in the audio ASR scheme. Besides, AVCDCN with audio-visual ASR still outperforms AVCDCN with audio-only ASR. On the other hand, the performances of audio-visual ASR without enhancement, with CDCN and with AVCDCN become very similar.

5. CONCLUSION AND PERSPECTIVES

In this paper, we investigate the use of the visual modality as a means to enhance noisy speech for robust speech recognition. We introduce a new non linear approach called AVCDCN that generalizes the existing audio-only technique CDCN. In the CDCN framework, a compensation vector computed as the weighted average of a set of pre-defined compensation codewords is added to each frame of noisy speech. In AVCDCN, the visual features are combined with the audio features to more accurately estimate the weight (the posterior probability) assigned to each of the compensation codewords. In our experiments on a multi-speaker digit task, AVCDCN provides significant performance gains over CDCN in both audio and audio-visual ASR schemes. In this paper, we report only on experiments where the noise added to the test data matches the noise seen during training. It would be interesting to consider situations where

AVCDCN and CDCN would be trained over a variety of different noises and SNRs, and, where the noise corrupting the test data would not necessarily match any of the noises seen during training. This would allow us to investigate the extent to which the visual modality, which is not affected by the specific characteristics of the noise, can make CDCN more robust to unseen noises.

6. REFERENCES

- [1] A. Acero and R.M. Stern. Environmental robustness in automatic speech recognition. *Proceedings of ICASSP'90*, pages 849–852, 1990.
- [2] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. Huang. High-performance robust speech recognition using stereo training data. *Proceedings of ICASSP'01*, pages 301–304, 2001.
- [3] L. Girin, J.L. Schwartz, and G. Feng. Audio-visual enhancement of speech in noise. *J. Acoust. Soc. America*, 6(109):3007–3020, 2001.
- [4] R. Goecke, G. Potamianos, and C. Neti. Noisy audio feature enhancement using audio-visual speech data. *Proceedings of ICASSP'02, in Press*, 2002.
- [5] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition, final workshop report, Center for Language and Speech Processing., 2000.
- [6] G. Potamianos, C. Neti, and J. Luetttin. Hierarchical discriminant features for audio-visual LVCSR. *Proceedings of ICASSP'01*, pages 165–168, 2001.