Lip Reading for Robust Speech Recognition on Embedded Devices

Jesús F. Guitarte Pérez [1,3] *, Alejandro F. Frangi* [2] *Eduardo Lleida Solano* [3] *and Klaus Lukas* [1]

[1] Siemens AG, Corporate Technology, Otto-Hahn-Ring 6, Munich, Germany.

[2] Department of Technology, Pompeu Fabra University, Pg Circumvallacio 8, Barcelona, Spain.

[3] Aragon Institute of Engineering Research, University of Zaragoza, María de Luna 1, Zaragoza, Spain.

jesus.guitarte@siemens.com, alejandro.frangi@upf.edu, lleida@unizar.es, lukas@siemens.com

Abstract

In this article a complete audio-visual speech recognition system suitable for embedded devices is presented. As visual feature extraction algorithms Active Shape Models (ASM) and Discrete Cosine transformation (DCT) have been investigated and discussed for an embedded implementation. The audio-visual information integration has also been designed by taking into account device limitations.

It is well known that the use of visual cues improves the recognition results especially in scenarios with high level of acoustical noise. We wanted to compare the performance of Lip Reading and the conventional Noise Reduction systems in these degraded scenarios, as well as the combination of both kinds of solutions. Important improvements are obtained especially for non-stationary background noises like voice interference, car accelerations or indicators clicks. For this kind of noises Lip Reading outperforms the results obtained with conventional Noise Reduction technologies.

1. Introduction

In recent years Automatic Speech Recognition (ASR) has been deployed widely in mobile phones and car environments due to convenience and safety reasons. However, especially in these scenarios often severe noise appear and has a very bad impact on the recognition rate. Several techniques like Noise Reduction based on acoustic signal processing [1] have been developed to improve the robustness of ASR when the acoustic signal is corrupted.

With the emerging distribution of cameras in embedded devices like mobile phones a new input modality is available: the visual information. Lip reading is a well known technique that exploits the additional information contained in the lips movement during speech [2] in order to improve the recognition rate and provide a more noise robust speech recognition system. Compared to conventional acoustic recognition, audio-visual speech recognition systems can decrease the Word Error Rate for various signal/noise conditions as it was achieve by Intel and IBM [3] [4]. These proposed solutions work on PC platforms, but are not specially designed to work on resource limited embedded devices. The challenge of this paper is to show that lip reading techniques can improve the recognition rate not only on PC but also with algorithms designed to work on embedded environments. Further improvements can be obtained in combination with conventional Noise Reduction systems.

This paper is organized as follows; in section 2 our Lip Reading System for embedded devices is presented. In section 3 the recognition results are summarized. Finally conclusions will be given in section 4.

2. Lip Reading System

Our Lip Reading System is made up of the following function blocks: the audio pre-processing on the acoustic channel, a lip localization system followed by a visual feature extraction on the visual channel and finally the integration of audio and visual information together with the recognition process.

In our implementation the audio pre-processing is going to be the same as used in conventional speech recognition systems for embedded devices [5]. In the following paragraphs the different blocks are going to be explained.

2.1 Lip Finding and Tracking

The first task to be solved in a lip reading system consists in the automatic detection and tracking of the mouth region. Our algorithm [6] is made up of two different functions: lip finding, and lip tracking. Lip Finding is applied when no previous information of the lip position is available. This happens in the first frame of a sequence or whenever the lips have not been correctly located in the previous frame. Lip Finding is based on a geometric model of the face. Structures of pixels are evaluated in order to know if their relative positions match a simplified prior model of the face, see figure 1.a. In particular, this model

accounts only for the relationships between location of the eyebrow(s) and the mouth.

Fig. 1.a: lip finding

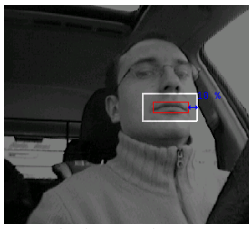
Fig. 1.b: lip tracking

Lip tracking proceeds when knowledge of the lips position is available in the previous frame. In this case we rely on the hypothesis that the position of the lips will not be very different between one frame and the next one. Lips will be searched in an area that is 10% larger than the region where the lips were located in the previous frame, see figure 1.b. Furthermore, lip tracking is more reliable and requires less resource than lip finding.
Our embeddable lip finding and tracking algorithm [6] is able to work without special light conditions as well as without any kind of reflected markers or special make up placed on speaker's lips.

2.2 Visual Feature Extraction

Once the mouth region is found an appropriate set of lip features must be extracted. The different feature extractions approaches that can be found in the literature have been classified according to the type of information source they process [4], *Shape Based* and *Appearance Based*. For this work one technique of each group was chosen; Active Shape Models (ASM) as *Shape Based* approach and Discrete Cosine Transformation (DCT) as *Appearance Based* one.

2.2.1 Shape Based Feature Extraction

In ASM a priori knowledge of the plausible mouth deformations is learnt in a training process. A set of points (landmarks) must be consistently located in the mouth contours of the training set. Rigid transformation dependencies are first removed by using Procustes Analysis. Then Principal Component Analysis (PCA) is applied on the aligned points. PCA computes the main variation modes of the points. This allows the deformations to be described only by a small set of parameters.
The eigenvectors ϕ_i and the associated eigenvalues λ_i of the landmark coordinates covariance matrix $\boldsymbol{S}$ are computed and sorted so that $\lambda_i > \lambda_{i+1}$. If Φ_i contains the t eigenvectors corresponding to the largest eigenvalues, a set of points describing the mouth contour x can be approximated by:

$$x = \bar{x} + \Phi \cdot b \qquad (1)$$

where $\Phi = (\phi_1 \mid \phi_2 \mid \ldots \mid \phi_t,)$ and $\boldsymbol{b}$ is a t-dimensional vector given by:

$$b = \Phi^T \cdot (x - \bar{x}) \qquad (2)$$

The $\boldsymbol{b}$ coefficients will describe the different variations of the mouth with respect to its mean value. A detailed description of the ASM with optimal features can be found in [7].
In our implementation the rigid transformation parameters, scaling, orientation and translation, obtained from the lip localization algorithm are used to simplify the matching of the landmarks. In this solution a horizontal filter with different polarity is used to track down the upper and the lower lip, which will define two regions. On the boundaries of these regions a first approximation of the lip contours is placed, afterwards PCA will be applied. For each subsequent image the new contours will be located using as initialization for the matching the contours points of the previous frame.

2.2.2 Appearance Based Feature Extraction

Appearance based methods provide visual features by using a grey scale intensity lip image transformation. These features contain information about the lip structure but also about the teeth and tongue visibility. In our system a cascade of operations [3] is performed before making the DCT. The coordinates of lip corners obtained by our lip finding and tracking algorithm [6] are used to determine the rotation, scaling and translation of the mouth. A new grey scale intensity lip image is obtained in such a way that the new normalized mouth is generated. Over this image a bi-dimensional elliptical and Gaussian mask is applied. Finally the image is filtered by a Gaussian filter in order to reject high frequency noises (see fig. 2).

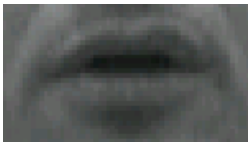
Fig. 2.a

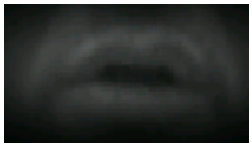
Fig. 2.b

Fig. 2: Mouth region before (2.a) and after (2.b) normalization and masking.

First and second DCT coefficients derivatives have been also used as features. The number of coefficients is quite large; a feature dimension reduction is achieved by a Linear Discriminate Analysis (LDA). This operation is quite simple since it comes down to a simple matrix multiplication. Moreover, LDA is an already implemented algorithm in conventional speaker independent speech recognition systems for embedded devices.

2.3 Audio-Visual Integration

Audio-visual integration solutions can be classified in three different groups: early integration, late integration and finally hybrid integration [4]. In all of these approaches the Hidden Markov Models (HMM) can be used to perform the recognition.
In hybrid integration the state emission probabilities from the HMM theory [8] are evaluated independently for the visual and the acoustic channel. However, the Viterbi decoding is performed only once. The integration is made on the emission probabilities level. This kind of feature combination is known as multistream [9].

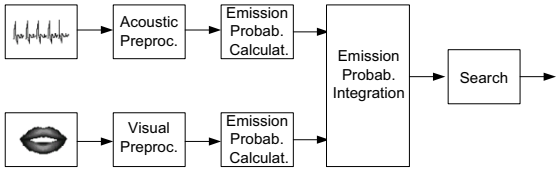

Fig. 3: Hybrid integration general diagram

Hybrid integration seems to be the most suitable solution for our implementation. On the one hand, the fact that two different emission probabilities are evaluated introduces a certain independency between the acoustic and visual channel. Furthermore, our hybrid integration delivers a more robust system compared to early integration. Both information sources can be weighted according to their respective reliability, e.g. signal to noise ratio for the acoustic channel is obtained and according to this one a fixed weighting is used, see table 1. On the other hand, performing only once the searching process (Viterbi decoding) saves a lot of resources, in comparison with the parallel search of late integration. This allows the audio-visual speech recognition system to be implemented in embedded devices without a substantial increase in CPU demands.

	SNR < 0 (dB)	*0 < SNR < 5 (dB)*	*5 < SNR < 16 (dB)*	*16 < SNR (dB)*
Weights	*V = 50%* *A = 50%*	*V = 20%* *A = 80%*	*V = 10%* *A = 90%*	*V = 0%* *A = 100%*

Table 1: Optimal Audio and Video weightings according to different SNR of the acoustic channel.

3. Results

All experiments shown in this article have been performed by using the CUAVE database [11]. This audio-visual database is composed by 36 American English speakers. 20 persons were used for training and other 16 for testing in order to obtain the speaker independent system. The experiments were always continuous digits "zero"-"nine" 4 times for every speaker making a total of 640 test numbers.

First of all DCT and ASM feature extraction will be compared. As it can be seen in table 2 the Word Error Rate using the DCT is 11.8% lower as using the ASM. For the ASM the detailed lip contours must be found, while for the DCT implementation only an approximation of the mouth region is required, which is much easier and works better in our embedded implementation. Due to the resource limitations, our ASM implementation performs only one iteration, which results in a sensitiveness to shadows and light conditions. ASM is often not able to find the lip contours properly, which explains the better results obtained by DCT. The results obtained with our DCT implementation improved the results obtained in [11] where WER of 63.2% was achieved also with DCT but with a lower resolution and without LDA. In our implementation a 64x32 pixel mouth region is analysed.

	ASM	*DCT*
Word Error Rate	*64.8 %*	*53.0%*

Table 2: Visual Feature Extraction Comparison for Continuous Speaker Independent digit recognition using only visual information.

In figure 4 the WER for different kind of background noises is shown. First of all a stationary noise is going to be examined. A car noise with constant speed from the Noisex database has been selected. As a second category of noise a mixture of stationary and non-stationary noises has been taken; a noise recorded in a car driving realistic situation. The motor noise is not any longer completely stationary due to accelerations and there appear other artefacts like indicator clicks. Finally, for non-stationary background noise a single speaker interference from the Macrophone database has been used. For these three kinds of noises we compare the results in figure 4 by using the two most common Noise Reduction solutions, Spectral Subtraction and Wiener Filter [1], and the use of our Lip Reading system for very degraded noise environments. As we can see, for a stationary noise, the conventional Spectral Subtraction is able to reduce the WER until 8%, outperforming Lip Reading. But when the noise is not stationary at all, as for example interfering voice, the use of Spectral Subtraction or Wiener Filter provides bad results with many insertions. Here, Lip Reading improves the WER significantly, as visual cues are independent on the acoustical noise. Even for the non stationary car noise Lip Reading outperforms the other two solutions.
Finally we have combined our visual cues with the acoustical signal improved by the conventional Noise Reduction techniques. For this experiment we have chosen the non-stationary car noise.

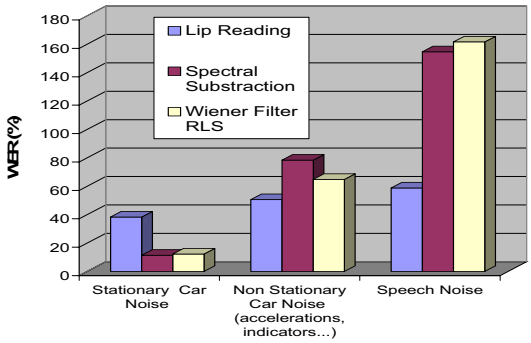

Fig. 4: Lip Reading and Noise Reduction conventional solutions (Spectral Subtraction and Wiener Filter) for different kind of noises and SNR = -10 dB.

In Figure 5 the WER for the different systems and combinations is shown for different SNR. As we can see for very bad SNR (-15 dB), Lip Reading reduces the WER obtained from Wiener Filter from 81% to 51%. But also for better SNR (5 dB) the combination of Lip Reading with Wiener Filter gives the lowest WER, improving the results of Wiener Filter in 24% relative.

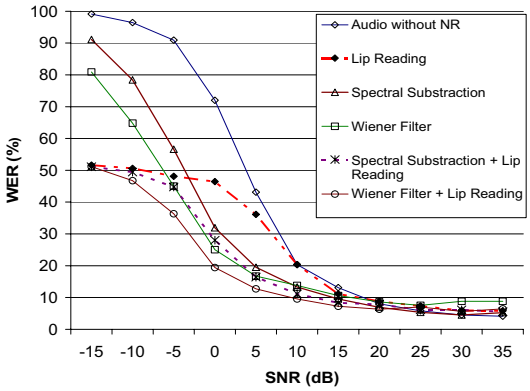

Figure 5: WER for different SNR with non-stationary car noise (accelerations, indicators clicks…)

4. Conclusions

It was shown, that significant improvements for degraded scenarios can be obtained even with a low resource Lip Reading System. For our resource saving implementation the DCT outperforms the results obtained by ASM, as it is a feature extraction that does not require a precise contour localization, but only an approximation of the mouth region.

Visual cues are independent of the kind of noise, which makes lip reading an interesting solution to improve the WER for non-stationary noises where the conventional Noise Reduction techniques fail.

Finally a combination of Lip Reading with conventional Wiener Filter outperforms all other investigated solutions for degraded as well as for more silent noise conditions.

5. Acknowledgements

The authors wish to express special thanks to T. Schneider, A. Schröer, N. Paragios and V. Ramesh for their good advises and helps. Siemens AG has supported the work of J. F. Guitarte under a PhD. contract. We would like to thank F. Althoff and S. Hoch from BMW Research and Technology for their support and helpful discussions within our joint project. A. F. Frangi is supported by a Ramón y Cajal Fellowship and by grant TIC2002-04495-C02 from the Spanish Ministry of Education & Science. We would like to thank the Clemson University for providing their audio-visual Cuave database.

6. References

[1] R. Singh, R. M. Stern, and B. Raj, "Signal and Feature Compensation Methods for Robust Speech Recognition," *CRC Press LLC, pp. 219-243, 2002.*

[2] H. McGurk, and J. MacDonald, "Hearing lips and seeing voices," *Nature, 264, pp. 746-748, December 1976.*

[3] A. Nefian, L. H. Liang, L. Xiao, X. X. Liu, X. Pi, and K. Murphy, "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP Journal of Applied Signal Processing, No. 11, pp. 1274-1288, 2002.*

[4] G. Potamianos, C. Neti G. Gravier, and A. Garg, "Recent advances in the Automatic Recognition of Audio-Visual Speech," *Proc. of the IEEE, vol. 91, No. 9, 2003.*

[5] I. Varga, S. Aalburg, B. Andrassy, S. Astrov, J. G. Bauer, C. Beaugeant, C. Geissler, and H. Höge, "ASR in mobile phones - an industrial approach," IEEE Trans. on Speech and Audio Processing, vol. 10, pp. 562-569, 2002.

[6] J. F. Guitarte, K. Lukas, A. F. Frangi, "Low Resource Lip Finding and Tracking Algorithm for Embedded Devices," *Proc. Eurospeech, Geneve, Switzerland, vol. 3, pp. 2253-2256, 2003.*

[7] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. ter Haar Romeny, and M. A. Viergever , "Active Shape Model Segmentation with optimal Features," *IEEE Trans. on Medical Imaging, vol. 21, No. 8, 2002.*

[8] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Procedings of IEEE, vol. 77, No. 2 pp. 257-286, 1989.*

[9] S. Dupont, and J. Luettin, "Audio-Visual Speech Modeling for Continuous Speech Recognition," *IEEE Trans. on Multimedia, vol. 2, No. 3, 2000.*

[10] S. Amarnag, S. Gurbuz, E. Patterson, and J. N. Gowdy, "Audio-Visual Speech Integration using Coupled Hidden Markov Models for Continuous Speech Recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing, 2003.*

[11] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new Audio-Visual Database for multimodal Human Computer Interaction Research," *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2002.*