Embedded Lip Reading for Automotive Environments

J.F. Guitarte Pérez [1,3] *, K. Lukas* [1] *,F. Althoff* [2] *, S. Hoch* [2] *and E. Lleida Solano* [3]

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

[2] BMW Research and Technology, Hanauer Str. 46, 80992 Munich, Germany.

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

jesus.guitarte@siemens.com, lukas@siemens.com, frank.althoff@bmw.de, stefan.hoch@bmw.de, lleida@unizar.es

Abstract

In this article a complete audio-visual speech recognition system suitable for embedded platforms is presented. As visual feature extraction algorithm Discrete Cosine transformation (DCT) has been selected for performance and robustness reasons. The audio-visual information integration has also been designed by taking into account device limitations.

Speaker dependency has been studied. There is an important improvement of the visual recognition rate by using speaker dependent systems, which implies that adaptation can be an important feature for future audio-visual recognition systems.

Car environment is normally degraded by many non-stationary 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 car environments due to convenience and safety reasons. However, especially in these scenarios often severe noise appear and have a very bad impact on the recognition rate. Several acoustic signal processing techniques like noise reduction have been developed to improve the robustness of ASR. With the emerging distribution of cameras, a new input modality is available to exploit visual information for a more noise robust speech recognition. Compared to conventional acoustic recognition, audio-visual speech recognition systems can decrease the Word Error Rate for various signal/noise conditions on PC.

The challenge of this paper is to show that lip reading techniques can improve the recognition rate even with algorithms designed to work on constrained embedded systems and that have to cope with difficult environments like automotive scenarios. Comparisons with classical noise reduction [1] systems applied to typical car noises will give hints for improvements by lip reading systems.

2. Embedded Lip Reading System

Our Lip Reading System is made up of the following function blocks as can be seen on figure 1. 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 [2].

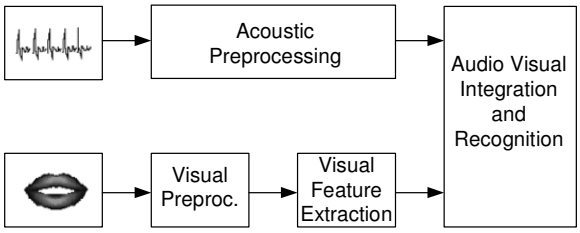

Fig. 1: General diagram of our Audio-visual speech recognition system for embedded devices

Our mouth detection algorithm [3] is made up of two different functions: lip finding, and lip tracking. 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. In particular, this model accounts only for the relationships between location of the eyebrow(s) and the mouth. Lip tracking proceeds when knowledge of the lips position is available in the previous frame. Our embeddable lip finding and tracking algorithm 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.

The coordinates of lip corners obtained by the previous algorithm 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 a new normalized mouth representation is generated. Over this image a bi-dimensional elliptical Gaussian mask is applied. Then a bi-dimensional DCT is performed and the

coefficients showed in figure 2 are selected as features for the recognition process.

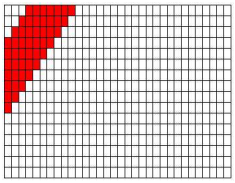

Fig. 2.a

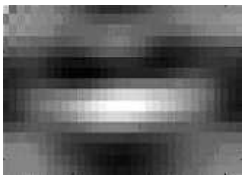

Fig. 2.b

Fig. 2: Selected DCT coefficients (2.a), inverse DCT of the selected coefficients (2.b)

First and second DCT coefficient derivatives have also been used as features. The resulting number of coefficients is quite large, thus a feature dimension reduction is achieved by a Linear Discriminate Analysis (LDA).
Audio and visual features are combined by using the well know multi-stream approach. In this integration strategy the state emission probabilities from the HMM theory are evaluated independently for the visual and the acoustic channel. A Viterbi decoding is performed only once in such a way that the integration is made on the emission probabilities level. This solution is the most appropriate for an embedded implementation because the integration on the emission probabilities level allows a dynamic weighting for both channels. Furthermore only one Viterbi decoding algorithm has to be performed.

3. Experiments and Results

First of all we have compared the results obtained by using a speaker dependent and a speaker independent system for the visual recognition. As we can see in table 1, there is an improvement of almost 15% of the WER when the system is trained for the user (speaker dependent). Therefore speaker adaptation will improve the results of the visual recognition and these visual features can be seen as a set of new features that are only added in the adaptation process.

	SD	*SI*
Word Error Rate	*37.0%*	*51.9%*

Table 2: Word Error Rate for speaker dependent (SD) and speaker independent (SI) recognition

In figure 3 speaker independent recognition results for different SNR are shown. For this experiment the CUAVE database [4] has been used. 20 persons were used for training and other 16 for testing. The experiments were always continuous digits "zero"-"nine" 4 times for every speaker making a total of 640 test numbers. An improvement by the usage of visual features in combination with the acoustic ones can be seen for SNRs lower than 10 dB. For better conditions acoustic information is more relevant. A noise recorded in a realistic car driving situation has been used as additive noise.

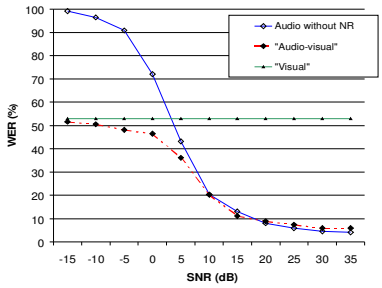

Fig. 3: WER for different SNR with non-stationary car noise (accelerations, indicators clicks…)

Finally we have performed experiments using two conventional acoustic noise reduction technologies, spectral subtraction and Wiener Filter. Moreover, a combination of these strategies together with the lip reading system has been investigated. As we can see, the best results are obtained by reducing the noise of the acoustical channel with the Wiener Filter and then combine this information with the visual channel.

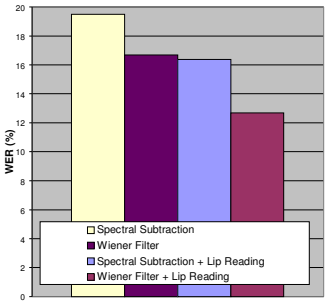

Fig. 4: Word Error Rate for different kind of acoustic noise reduction strategies and the combination with visual information for SNR = 5 dB

4. 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] 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.*

[3] 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.*

[4] 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.*