AudioVisual Speech Recognition Using Motion Based Lipreading

María José Sánchez Martínez, Juan Pablo de la Cruz Gutiérrez

Infineon Technologies AG
Corporate Research System Technology
Munich, Germany

maria.sanchez@infineon.com JuanPablo.delaCruz-Guiiterrez@infineon.com

Abstract

This paper presents an audio-visual speaker-dependent continuous speech recognition system. The idea is to extract features from the audio- and the video-stream of a speaking person separately and use them to train a Hidden-Markov-Model based recognizer with the combined feature vectors. While the audio feature extraction follows a classical approach, the visual features are obtained by means of an advanced image processing algorithm which tracks certain regions on the speaker’s lips with high robustness and accuracy. For a self-generated audio-visual database, we compare the recognition rates of audio only, video only and audio-visual based recognition systems. We also compare the results of the audio only and the audiovisual systems under different noise conditions. The work is part of a larger project which aims at a new man-machine interface in the form of a so-called Virtual Personal Assistant which communicates with the user based on the multimodal integration of natural communication channels.

1. Introduction

In human-to-human dialogs, the partners integrate many different communication channels, such as speech, gaze and mimics to increase the robustness of the information transfer. In particular the recognition of spoken language in presence of background noise is greatly enhanced by observing the speaker’s lip movements simultaneously (“lip-reading”). The audio channel and the video channel differ in many respects and in addition, there is usually a considerable temporal delay between the vision- and audio signal. However, the human brain is capable of extracting the important features from both channels, to integrate them and to recognize the spoken words. Audio-visual speech recognition deals with the transfer of this capability into man-machine-communication interfaces.

The paper on hand describes our approach to an audio-visual speech recognition module for an ambitious man-machine-interaction scenario. This scenario, which will be described in detail in the next section, aims at a multi-modal communication interface which consists of speech-recognition and -output, the recognition and control of head-, hand- and body gestures and an advanced scene interpretation module. The idea is to integrate as many different communication channels as possible to mimic the robustness of the human-to-human dialog. While we have implemented already many of the recognition modules separately (e.g. pointing- and head-gesture-recognition), the fusion of audio and lip-reading is our first step towards an integration of natural communication channels. As our main intention is to investigate the potential benefits of integration, we decided to use a well known, classical speech recognition system, namely the Hidden-Markov-Model (HMM) recognizer HTK by the University of Cambridge [1] and a lip-tracking algorithm we developed on our own. Our idea is to extract characteristic high level visual features from the lip movements, to express them as time series of the same format as the audio features and to provide a concatenated feature vector stream to the HMM for the recognition process.

In the following, we will describe this approach and the results step by step. First, we briefly present the concrete man-machine interaction scenario. Next, we describe our audio-visual database which we use to train the recognizer with. Further, we explain our implementation of the HTK and the image processing algorithm that underlies the lip-tracking mechanism. Finally, we will present the results that we achieved in varying background noise conditions.

2. The Virtual Shop scenario

As a concrete application for the man-machine-interface we have selected an E-commerce scenario that we call a virtual shop. Herein, a user purchases items (in our case furniture) from an online store. The user communicates with a virtual personal assistant (VPA) in the form of a 3D animated talking head. This assistant presents the items, gives information on the ones selected, and asks the user whether he wants to purchase them. The user can answer the assistant’s questions by means of gestures, speech, etc.).

Figure 1: The Virtual-Shop scenario: a user communicates with a virtual personal assistant (VPA) on the screen (lower right) by means of natural communication channels (pointing, gaze, speech, etc.). The VPA describes items (upper left) which the user can purchase. The final application will also incorporate the possibility to communicate with a human partner via video conference (upper right).
lected by the user and reacts on simple speech commands that are provided by the audio-visual recognizer. In addition, the VPA reacts on head and hand gestures, mimics and gaze direction captured by means of a simple web-cam. The behavior of the VPA is controlled by an advanced behavioral organization mechanism which integrates all available sensor inputs and generates an internal representation of the scene. This scenario determines the task domain of the speech recognizer; the vocabulary to be recognized consists of typical phrases appearing in a sales conversation.

3. Audiovisual Database

Current freely available audio-visual databases provide only sequences of spoken digits as train/test sets. However, in the virtual shop scenario the recognition of keywords and full sentences is required. Therefore, we generated our own audio-visual database containing videos from a single male person speaking the 67 sample sentences from the TIMIT[2] corpus as training data. For the test data set, we recorded 12 phrases (7 words per sentence on average) related to the aforementioned sales conversation. The videos were recorded as 720x576 pixel, 25 fps, DV-coded stream using a digital Panasonic® NV-EX1 DV-Camera. The corresponding audio channel was recorded as 44.1 kHz mono PCM-coded stream using a Vivanco® DV-Camera. The audio and the video channels were fed into an IBM® X30 notebook and were multiplexed in real time into uncompressed avi-video files using the Microsoft® DirectShow® program GraphEdit. Later, these videos were resized to 320x240 pixels and compressed using the DivX® 5-codec and the freeware video editor VirtualDub.

To investigate the general feasibility of our audio-visual recognition approach, we work on this audio-visual database as a defined test-bed: from the training- and test-videos, audio features and video features are extracted as described in the following section. The several test runs with varying audio-noise levels described in the results-section are run off-line always with the same corpus to ensure comparable conditions.

4. Audio-Visual Speech Recognition Systems

The two major issues to be considered in the design of an audio-visual speech recognition system besides the audio-only recognizer are the design of a visual front-end where visual speech features are accurately and reliably extracted and the development of an appropriate strategy to integrate the two separate information sources. In the design of the visual front-end, most well-known approaches to visual feature extraction are AAM, ASM, MSA [3] and image transforms, like DCT [4]. They all proceed in three basic steps. First, image landmarks (as in ASM and AAM) or the region of interest (ROI, as in DCT and MSA approach) containing the speaker’s mouth area, are labelled. Then, the visual components are extracted, accounting for a) high-level features (lip shape estimate in ASM), b) low-level features built upon pixel values (scale-space decomposition in MSA, or DCT coefficients) or c) a combination of both (AAM). Further, postprocessing is usually applied to produce a smaller subset of more robust features. In order to incorporate the dynamics of information, visual features from consecutive frames are concatenated. A more recent algorithm based on image transform is that developed by Potamianos and collaborators [5].

In our approach, on the contrary, visual feature extraction is performed on the basis of an accurate motion tracking algorithm based on optical flow, as described in [6].

The fusion strategies of both information streams can be grouped into two main categories: feature fusion and decision fusion methods [7]. In the first method, a single recognition system is trained on the concatenated vectors of audio- and visual features. Plain feature concatenation, feature weighting, hierarchical discriminant feature extraction, and dominant and motor recording fusion methods belong to this fusion category. The second method can capture the reliability of each information stream by appropriately combining the likelihoods of single-modality HMM recognizers’ decisions. This likelihood recombination can occur at various levels of integration: state level (giving rise to a multi-stream HMM), phone level (that can be modelled by means of the product or coupled HMM) and utterance level (based on a discriminative model combination approach and rescoring of n-best hypotheses).

In this work, we use the plain feature concatenation method which turns out to be already very successful in performing the integration of the audio and visual streams.

5. Audio & Visual Features

The first stage in the implementation of the recognition system is the encoding of the corresponding audio- and visual information streams into appropriate feature vectors.

5.1. Audio Features

For encoding the audio features, we follow a rather conventional approach, which we will therefore describe only briefly: the audio signal is sampled at a frame period of 10ms. Each frame is then coded using a 39 dimensional vector consisting of 12 Mel Frequency Cepstral Coefficients with one energy parameter and their first and second derivatives.

5.2. Visual Features

As stated before, a motion tracking algorithm is applied to the user’s face images to get the visual features. Based on results from psychophysical experiments, like those from studies of the McGurk effect [8], we measure the motion of facial elements that are relevant for the perception of speech. These elements are the mouth corners, the mid of lower and upper lip, nose and chin. The image regions of these elements (regions of interest, ROIs) are initially marked manually (see Figure 2). An estimate of the displacement of the central point of each ROI, and its variance, is calculated from every two consecutive frames.

As we work off-line anyway, this manual lip-finding does not pose a problem for our general feasibility investigation. However, for the practical application in the Virtual-Shop scenario, the user’s lips must be detected automatically. Currently, we develop a visual recognition algorithm which identifies the facial elements based on the face position. We can already find the position of the face based on skin-color
5.3. Building Audio-Visual Feature Vectors

The fusion of the audio- and visual information streams is made at feature level. The joint audio-visual feature vector is obtained by the synchronous concatenation of the audio- and visual feature vector. In the database audio and video was recorded simultaneously. The audio feature vector is sampled at 100 Hz from the audio signal. The video has a frame rate of 25 Hz and for every frame, the positions of the facial elements are determined. That means that each frame in the video signal corresponds to four samples in the audio feature vector. Therefore, we upsample the video signal by a factor of four. Then, we obtain the audio-visual feature vector by just concatenating the two feature vectors of audio and video 100 times per second. This process results in a 44 dimensional audiovisual feature vector which we provide as input for the Hidden-Markov-Model (HMM) based HTK-speech-recognizer.

6. Baseline System

Using the HTK toolkit we build a baseline system which can be trained on audio-only, visual-only or on audio-visual features.

The HMMs are continuous density mixture Gaussian tied-state triphones.

First we built monophone models with single Gaussian densities. A set of 50 phonemes was used, each of them modelled by a 3-state left-right model. These models are initialized with identical values of the mean and variance, and then retrained. A short pause model (which has only one state) is added and tied to the center state of the silence models and two retrain passes are run. A forced alignment is performed to get the pronunciations that best match the acoustic data when the dictionary contains different pronunciations of the words. Once these new transcriptions are retrained, the monophone models are already trained. Context dependent triphone HMMs are then created using phonetic decision trees. After two training iterations, we get the trained triphone models.

7. Experimental Results

To measure the recognition results of the audio-visual recognizer, we compare it with the results for an audio-only and a vision-only system. For each modality, we implemented a HMM-based recognition system. Training and decoding of the models were performed using the HTK toolkit. We built one recognizer for each modality: audio only, visual only, and audio-visual. For the audio- and the audio-visual modalities, the models were trained using clean audio. The tests were then performed under matched conditions, that is, also clean audio was used for testing. Table 1 shows the results for this test-condition:

<table>
<thead>
<tr>
<th>System</th>
<th>Audio</th>
<th>Visual</th>
<th>Audio-Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>WER(%)</td>
<td>2.02</td>
<td>36.67</td>
<td>0</td>
</tr>
</tbody>
</table>

Next, we measured the recognition results of the audio only, and the audio-visual recognizers when tested in noisy conditions. We added artificial white noise at different Signal to noise ratios (SNR) to the clean test data to simulate various noisy test conditions.

Given the desired value of the SNR and the power of the clean audio signal, the power of the noise signal to be added is calculated by applying

\[ \text{SNR} \text{ (dB)} = 10 \log \left( \frac{\sigma^2}{\sigma_n^2} \right) \]  

where \( \sigma^2 \) is the power of the clean audio signal and \( \sigma_n^2 \) is the power of the noise signal.

Using this procedure, we generated test audio data at 10, 15, and 20 dB and we conducted different experiments under mismatched conditions, that is, the training was made on clean data in all cases. Figure 3 shows the results of the experiments. For each value of the SNR, a difference of proportions significance test was applied to determine the statistical significance of the differences in the audio- and audio-visual WERs (see table 2 and figure 3).

These results show that integrating the visual information channel enhances speech recognition in particular in the pres-
Figure 3: Word Error Rates (WER, in percent) of the audio, visual and the audio-visual recognition system under various signal-to-noise ratios. Artificial white noise was added. For noisy data, the tests were conducted under unmatched conditions.

Table 2: Word Error Rates of the audio and the audio-visual recognition systems at various SNRs under unmatched conditions. Differences in WERs and level of significance of those differences (calculated using a difference of proportions significance test).

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>Audio WER(%)</th>
<th>Audio Visual WER(%)</th>
<th>Δ WER (%)</th>
<th>Significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>57.78</td>
<td>27.78</td>
<td>30</td>
<td>0.0001</td>
</tr>
<tr>
<td>15</td>
<td>41.11</td>
<td>11.11</td>
<td>30</td>
<td>0.0001</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>5.56</td>
<td>4.44</td>
<td>0.2</td>
</tr>
<tr>
<td>clean</td>
<td>2.02</td>
<td>0</td>
<td>2.02</td>
<td>0.1</td>
</tr>
</tbody>
</table>

ference of high noise levels. It is also interesting to observe, that although the visual channel alone has a relatively low recognition rate, it can raise the audio-visual recognition rate in all test conditions. Obviously, our simple method of concatenating the feature vectors is already sufficient for the HMM to extract that portion of additional information from the vision channel that is needed to improve the recognition rate of the audio-visual system.

8. Conclusions

Aiming at a man-machine-interaction scenario which is based on the integration of multiple natural communication channels, we have developed a mechanism for fusing audio- and vision data for improving speech recognition. The idea is to exploit the inherent redundancy of the two modalities audition and lip-reading to enhance the robustness of the recognition. By making use of the conventional speech recognition toolbox HTK and an optical flow based image processing algorithm, we obtained two feature vectors which we concatenated and streamed into a HMM-based speech recognition system. The results of comparison tests under various noise conditions revealed that the audio-visual recognition system reaches a higher recognition rate than audio-only or video only. In particular at high noise levels, adding the visual information can improve the recognition rate substantially. Of course, the training- and test-corpus was rather small and further studies with larger data bases and more test sentences will have to be conducted to ver-

ify the results. However, our aim was to build an audio-visual recognizer as simple and with as few data processing steps as possible. With a larger database, the recognition rate would most likely fall. However, we expect that we could still compensate for this introducing additional data processing algorithms (such as ICA, used by other, more sophisticated approaches) that would reduce even more the correlation between the information streams.

Taking into account the surprisingly simple fusion approach of just concatenating feature-information from the two very different modalities, supports our idea that there exists a common level on which interaction information can be integrated independently of the source of information. In our case this level consisted in audio and visual features that characterize speech. To find more evidence for this idea, the next step will be the integration of symbolic information in the style of late fusion: an example is a user who indicates an object by means of a pointing gesture in case that speech recognition alone can not resolve an ambiguous situation during the communication (e.g. "show me that thing!").

On the short time scale, we plan to incorporate a visual lip-detection mechanism to allow for an on-line audio-visual recognition without having to mark the initial position of the face elements manually.

9. References