Labeling Audio-Visual Speech Corpora and Training an ANN/HMM Audio-Visual Speech Recognition System

Martin Heckmann°, Frédéric Berthommier®, Christophe Savario®, Kristian Kroschel°

° Institut für Nachrichtentechnik
Universität Karlsruhe
Kaiserstraße 12, 76128 Karlsruhe, Germany
{heckmann, kroschel}@int.uni-karlsruhe.de

® Institut de la Communication Parlée (ICP)
Institut Nationale Polytechnique de Grenoble
46, Av. Félix Viallet, 38031 Grenoble, France
{bertho, savario}@icp.inpg.fr

ABSTRACT

We present a method to label an audio-visual database and to setup a system for audio-visual speech recognition based on a hybrid Artificial Neural Network/Hidden Markov Model (ANN/HMM) approach.

The multi-stage labeling process is presented on a new audio-visual database recorded at the Institute de la Communication Parlée (ICP). The database was generated via transposition of the audio database NUMBERS95. For the labeling first a large subset of NUMBERS95 is used to achieve a bootstrap training of an ANN, which can then be employed to label the audio part of the audio-visual database. This initial labeling is further improved via readapting the ANN to the new database and reperforming the labeling. From the audio labeling then the video labeling is derived.

Tests at different Signal to Noise Ratios (SNR) are performed to demonstrate the efficiency of the labeling process. Furthermore ways to incorporate information from a large audio database into the final audio-visual recognition system were investigated.

1 INTRODUCTION

Perceptual studies show that humans use both, the acoustic information and the speakers lips movement to recognize what was said [1][2]. This led to different approaches to integrate the visual information also in automatic speech recognition systems [3][4]. The integration of both data streams is most promising in very adverse acoustical conditions, where the recognition rate of audio recognizers drops significantly whereas the video path is not affected. The systems proposed so far are very different in many respects. There are large differences in the size of the database used, the task envisaged, as vowel recognition, word recognition or continuous word recognition, and the underlying system structure. Even though the size of the databases and the complexity of the tasks increased over the years, the databases used are still rather moderate and the tasks are rather simple compared to standard automatic speech recognition systems.

ANN/HMM hybrid models have shown to provide very good recognition results for automatic continuous word speech recognition [5]. We therefore apply this concept to audio-visual continuous word recognition. One major drawback of the ANN/HMM hybrid models is their need for extensive training data due to the huge amount of free parameters in the ANN’s. Especially for audio-visual speech recognition this is a strong constraint, as so far only audio-visual databases with a moderate size are available and the recording of new databases is very time consuming.

As a consequence of the poor availability of audio-visual databases we recorded a new database to setup our system. To simplify the recording of the database we copied the structure and part of the contents of the audio database NUMBERS95 (NB95) from the Oregon Graduate Institute (OGI) to our new audio-visual database. Hence we will refer to this new database as Audio-Visual NUMBERS95 (AVNB95). The database NUMBERS95 is easily available and experiences in setting up an hybrid ANN/HMM audio recognition system with this database were already gathered. Therefore this database is an ideal choice to base a new audio-visual database on it. The concept of multi-stage training introduced in this paper facilitates the labeling of the newly recorded audio-visual database. A complete automatic process, which does not rely on labeling by humans is presented.

2 SYSTEM STRUCTURE

We use an ANN/HMM hybrid model for continuous audio-visual word recognition. The implementation was carried out using the tool STRUT from TCTS lab Mons, Belgium [6]. We perform the labeling of the database with an audio-visual speech recognition system based on a Separate Identification (SI) structure (compare Fig. 1 and see [3] for more details).

Audio feature extraction is performed using RASTA-PLP [7] and the video features are extracted via a chroma key process, which requires coloring of the speakers lips with blue ink. Due to the coloring, the lips can then be located easily and their movement parameters can be extracted in real time. As lips parameters

- outer lip width
- inner lip width
3 SETUP OF THE DATABASE

In this section the acquisition and the labeling of the audio-visual database via the multi-stage training process are presented. Labeling was performed with a SI structure because it allows a separate treatment of the audio and video part, which is necessary for the multi-stage training process.

3.1 Database Acquisition

The audio-visual database to be labeled was newly recorded at the ICP. Selected utterances from NUMBERS95 were chosen and repeated by a native English-speaking male subject. Transposing parts of the original audio database NUMBERS95 to a new audio-visual database facilitates some key points in the setup of the database. First experiences were already made in setting up a hybrid ANN/HMM recognition system using NUMBERS95. The parameter values found there were helpful for establishing the audio visual system. Also the corresponding phrase models and the dictionary could be directly transferred. Furthermore it was not necessary to generate the text transcriptions manually, because they could be taken from NUMBERS95.

For the new database total of 1700 sentences or 6400 words were recorded. The database was subdivided into two subsets of equal size for training and final recognition. Synchronous recordings of the speech signal and video images of the head and mouth region at 25 frames per second were taken. As location a room with small reverberations and good screening to environmental noise was chosen to achieve a reasonable signal to noise ratio. Recordings were made on BETACAM video and standard audio tapes and A/D converted with 8kHz off-line. The video parameter were interpolated to 8kHz in order to be synchronous with the audio data.

3.2 Multi-Stage Labeling

Before using the newly recorded database for training, labeling is necessary. This means that for the training part of the database, for each speech frame, the actual uttered phoneme has to be determined. One common way to achieve the labeling is a complete manual labeling. Another is a manual labeling of a subset of the database, training of the ANN on the subset and using the ANN to label the rest of the database. And finally, the use of an ANN trained on a different database to label the newly recorded database. We extended the last labeling approach to the problem of labeling an audio-visual database (see Fig. 3). Special emphasis was set to avoid human interference in the labeling process and thus make it also applicable to large databases.

Pretraining on NUMBERS95

The first stage of our multi-stage training approach consists of pretraining the audio part of our recognition system on a large audio database. In our case of course NUMBERS95 was chosen. We have selected 3000 utterances from different speakers under varying conditions from the NUMBERS95 database and performed a cross-check during training on an independent test set to avoid over-adaption (compare [6]).
Figure 3: Steps performed during the multi-stage labeling of the audio-visual database

Segmentation of the audio channel of the database

After pretraining on NUMBERS95 the ANN can be used to perform the segmentation of the newly recorded database, provided that the recognition rate of the ANN on the audio-visual database is not too low. Otherwise too many errors occur during the segmentation. Comparison of the recognition rates of the ANN on NUMBERS95 and the audio-visual database (AVNB95) shows that the recognition rates are unfortunately rather low on AVNB95 (compare Tab. 1). This is due to the very different recording conditions of NUMBERS95 and AVNB95 and the fact that in one case recordings of different American speakers and in the other case of one English speaker were made. A first segmentation of AVNB95 is nevertheless feasible as not really each phoneme has to be recognized but only their time-alignment in the sentence has to be determined. This is because the succession of the single phonemes for each sentence can be derived from the text-transcription of the sentence and the dictionary which provides the sequence of phonemes for each word. To achieve the time-alignment, a so called forced Viterbi alignment is used. Here for each sentence the alignment of the phonemes which gives the highest overall likelihood is chosen. The likelihoods are calculated from the ANN.

<table>
<thead>
<tr>
<th>NB95</th>
<th>AVNB95</th>
<th>AVNB95</th>
<th>AVNB95</th>
<th>AVNB95</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>train-</td>
<td>AVNB95</td>
<td>AVNB95</td>
<td>AVNB95</td>
<td>AVNB95</td>
</tr>
<tr>
<td>ing on</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word Error Rate</td>
<td>11.6%</td>
<td>28.5%</td>
<td>7.1%</td>
<td>4.0%</td>
</tr>
</tbody>
</table>

Table 1: Recognition scores at different training steps

Relabeling the audio part

Following the segmentation, training of the audio part on the audio-visual database can be performed. The word error rates achieved with the new ANN trained on the audio part of the new audio-visual database are by far inferior to those resulting from using the ANN trained on NUMBERS95 (compare Tab. 1). The large improvement of the recognition scores opens the possibility to improve the first labeling of the database via reperforming the forced Viterbi alignment with the new ANN (compare Tab. 1). The resulting final audio labeling is then used for the video labeling. Due to the assumption that the corresponding acoustical and visual articulations represent the same phoneme, labeling of the video part of the database is directly derived from the audio labeling. This bootstrap video labeling can then be used in subsequent steps to achieve an improved labeling via the introduction of more complex viseme models which, for example, take into account that visemes are in fact only a subset of the phonemes.

To verify the video labeling of the database, an ANN was trained on the video part. The resulting word error rates can be seen in Tab. 1.

4 TRAINING AND TEST

The huge size of the ANNs used for phoneme identification (about 10^6 free parameters) requires extensive training data to obtain meaningful recognition results. Even though the audio-visual database available has a considerable size, its size is still rather moderate compared to the size of the ANNs. Therefore we also investigated ways to incorporate information from a large audio database into the final audio-visual system to improve recognition results. We compared the effects of pretraining the audio part of our system first on NUMBERS95 and then continuing the training on AVNB95, with the recognition scores from training the audio part directly on AVNB95. The intention was to see if a preliminary adaptation of the weights of the ANN via pretraining on NUMBERS95 will help to improve the final recognition task on AVNB95.

Furthermore we performed recognition tests on AVNB95 at different SNR levels with the SI system as presented in Sec. 2. These tests demonstrate the efficiency of the labeling process.

4.1 Pretraining the Audio Part

To evaluate the effects of the pretraining on the final recognition, we first trained the ANN of the audio part as described in Sec. ‘Pretraining on NUMBERS95’. In the next step we used the audio part of AVNB95 to continue the training. We compared this system with the one we trained directly on AVNB95. The results can be seen in Fig. 4. At high SNR an improvement of 1 – 2% due to the pretraining can be seen. Whereas with decreasing SNR the recognition scores achieved with the pretrained system drop significantly below those resulting from the system trained from scratch on AVNB95.

4.2 Recognition Tests with a SI Structure

In order to evaluate the results of the labeling of the database and to probe the potential of the hybrid ANN/HMM concept for audio visual speech recognition, we performed recognition tests
with a SI structure at different SNR levels. The SI system was trained directly on the new database AVNB95. For the tests white Gaussian noise was added to the database. The fusion process was performed as detailed in Sec. 2.

Fig. 5 compares the results of the audio and video only recognition with the audio part only.

We presented a new methodology to label an audio-visual database and to set up an audio-visual speech recognition system based on an hybrid ANN/HMM system. The labeling of the database was performed in multiple stages and enabled a fully automatic process. The labeling process was simplified by transposing a well known audio database, in our case NUMBERS95, to an audio-visual database. Recognition tests at different SNR levels were performed to show the good performance of the labeling process and to show the potential of a hybrid ANN/HMM system for audio-visual speech recognition. However, to achieve meaningful recognition results improvements of the system are necessary, e.g. a more complex viseme model and a more elaborated fusion strategy.

We further investigated the effects of integrating information from pretraining on a large audio database into the final system. The results showed, that for high SNR values, a small improvement can be achieved, whereas for low SNR’s the pretraining has a strong negative effect. This can be due to the very different recording conditions of our database and NUMBERS95, which was recorded over telephone lines. Nevertheless pretraining can be advantageous in the labeling phase of a database. When the results of the first labeling alone are not sufficient to train the ANN from scratch, recognition scores and thus the labeling can be improved by continuing the training of the pretrained ANN on the new database.

6 ACKNOWLEDGMENTS

We want to thank Alain Arnal for the recording of the audio-visual database and John Barker for his great patience while serving as the speaker. Furthermore we want to thank Thorsten Wild, Vincent Charbonneau and Christian Sørensen for carrying out most of the simulations.

This work was in part supported by the European TMR network SPHEAR.

REFERENCES