A Multi-stage Methodology to Setup an
ANN/HMM Audio-visual Speech
Recognition System

Martin Heckmann * Frédéric Berthommier **
Vincent Charbonneau * Kristian Kroschel *

* Institut für Nachrichtentechnik
Universität Karlsruhe
Kaiserstraße 12, 76128 Karlsruhe, Germany
{heckmann, kroschel}@int.uni-karlsruhe.de
vincent.charbonneau@enst-bretagne.fr

** Institut de la Communication Parlée (ICP)
Institut National Polytechnique de Grenoble
46, Ave Félix Viallet, 38031 Grenoble, France
bertho@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. Furthermore we introduce
a new audio-visual fusion method incorporating a priori probabilities.
Tests at different Signal to Noise Ratios (SNR) are performed to demonstrate the
efficiency of the labeling process and to evaluate the performance of the new fusion
modality.

Keywords: Speech recognition, Audio-visual database, Automatic labeling,
ANN/HMM, Audio-visual fusion

1. INTRODUCTION

Perceptual studies show that humans use both,
the acoustic information and the speakers lips
movement to recognize what was said (Sumby and
Pollack, 1954). This led to different approaches
to integrate the visual information also in auto-
matic speech recognition systems (Rogozan and
Délégue, 1998)(Teissier et al., 1999). The inte-
gration of both data streams is most promising
in very adverse acoustical conditions, where the
recognition rate of audio recognizers drops signifi-
cantly 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 pro-
vide very good recognition results for automatic
continuous word speech recognition (Morgan and
Boulard, 1995). 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 of 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 set up 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. 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 (of Mons University, 1997). 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 (Rogozan and Deléglise, 1998) for more details).

![Fig. 1. Separate Identification (SI) audio-visual speech recognition system](image)

Audio feature extraction is performed using RASTA-PLP (Hermansky et al., 1992) 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 were chosen.

To take temporal information into account, several time frames of the audio and video feature vectors are presented simultaneously at the input of the corresponding ANNs. The individual phonemes are modeled via left-right HMM models. The order of the HMMs used to represent the different phonemes was adapted to the mean length of the corresponding phoneme. Word models are generated by the concatenation of the corresponding phoneme models. Recognition is based on a dictionary with the phonetic transcription of 30 English numbers. Complete sentences containing a sequence of numbers were presented to the system during the recognition process. No distinction between phonemes and visemes was made in the fusion process. Each acoustical articulation is assumed to have a synchronously generated corresponding visual articulation. Even though this is a rather coarse approximation, it facilitates the labeling of the audio-visual database.

Fig. 2 shows the resulting ANNs. For the audio path 13 parameters over 9 time frames are extracted and their actual value and their first and second order derivative are forwarded to the ANN. In the video path the 6 parameters of the lips evaluated in 9 time frames and their derivatives are used.

![Fig. 2. ANN used in the SI structure](image)

**Fusion of the audio and video path**

In the SI structure during the audio-visual fusion process the final a posteriori probability \( P(H_i|x_A, x_V) \) for the occurrence of the phoneme \( H_i \) is calculated from the two a posteriori probabilities \( P(H_i|x_A) \) and \( P(H_i|x_V) \) of the phoneme \( H_i \). These posteriors are estimated by the corresponding ANNs from the acoustic feature vector \( x_A \) and the visual feature vector \( x_V \), respectively. Assuming conditional independence of the visual and acoustical features as in (Movellan and Chaderdon, 1996):

\[
P(x_A, x_V|H_i) = P(x_A|H_i)P(x_V|H_i)
\]

this leads to:

\[
P(H_i|x_A, x_V) = \frac{P(H_i|x_A)P(H_i|x_V)}{P(H_i)}
\]
for the final posterioris after applying Bayes’ rule. The ANNs only supply an estimate \( P \) of the true a posteriori probabilities \( P \). Unfortunately, in the case of the audio part the quality of this estimate is strongly dependent on the SNR. Therefore the final a posteriori probability is calculated as a weighted product of the acoustic and visual a posteriori probabilities dependent on the SNR. For high noise levels the estimation should only depend on the estimate from the video path, thus \( \hat{P}(H_i|\mathbf{x}_A,\mathbf{x}_V) \rightarrow \hat{P}(H_i|\mathbf{x}_V) \) and for clean speech it should in some cases only depend on the audio estimate as the video estimate in general is less precise. Consequently \( \hat{P}(H_i|\mathbf{x}_A,\mathbf{x}_V) \rightarrow \hat{P}(H_i|\mathbf{x}_A) \). To take these postulations into account Eq. 2 is rewritten as:

\[
\hat{P}(H_i|\mathbf{x}_A,\mathbf{x}_V) = \frac{\hat{P}^\alpha(H_i|\mathbf{x}_A)\hat{P}^\beta(H_i|\mathbf{x}_V)}{\varepsilon'(\alpha, \beta)}
\]

(3)
The terms \( P(\mathbf{x}_A) \), \( P(\mathbf{x}_V) \) and \( P(\mathbf{x}_A,\mathbf{x}_V) \) do not depend on the phoneme \( H_i \) and can thus be combined to a normalization factor \( \varepsilon \). The normalization factor \( \varepsilon \) has to guarantee that \( \sum_{j=0}^{N-1} \hat{P}(H_j|\mathbf{x}_A,\mathbf{x}_V) = 1 \) and therefore is set to:

\[
\varepsilon'(\alpha, \beta) = \frac{1}{\sum_{j=0}^{N-1} \frac{\hat{P}^\alpha(H_j|\mathbf{x}_A)\hat{P}^\beta(H_j|\mathbf{x}_V)}{\varepsilon'(\alpha, \beta)}}
\]

(4)
The a priori probability \( \hat{P}(H_i) \) is estimated from the training set. The weighting factors \( \alpha \) and \( \beta \) depend both on a third parameter \( c \) according to:

\[
\alpha = \frac{1}{1 + \exp(-c - 5)} \quad \beta = \frac{1}{1 + \exp(c - 5)}
\]

(5)
where \( c \) is a function of the SNR, which is presented to the recognition system as contextual information. The additive constant in the exponents of Eq. 5 is set in order to obtain \( \alpha(0) = \beta(0) \approx 1 \). When only the audio or only the video path contributes to the final a posteriori probability \( (\alpha = 1, \beta = 0 \text{ or } \alpha = 0, \beta = 1) \) \( P(H_i)^{\alpha+\beta-1} = 1 \), leading to the desired marginal probabilities.

Eq. 3 is in contrast to:

\[
\hat{P}(H_i|\mathbf{x}_A,\mathbf{x}_V) = \frac{\hat{P}^\lambda H_i(\mathbf{x}_A)\hat{P}^{\lambda H_i}(\mathbf{x}_V)}{\delta(\lambda)}
\]

(6)
as used in (Teissier et al., 1999) and in a similar form in (Rogozan and Delégilse, 1998)(Adjoudani and Benoit, 1996) where \( \lambda \) is a weighting factor for the audio and video path and \( \delta(\lambda) = \sum_{j=0}^{N-1} \hat{P}^{\lambda H_j}(\mathbf{x}_A)\hat{P}^{\lambda H_j}(\mathbf{x}_V) \) Here the priors of the phoneme \( H_i \) are not taken into account for the calculation of the final a posteriori probability. Therefore we will compare in Sec. 4 our results to those achieved when applying Eq. 6.

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 phoneme 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 a 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. 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.

**Audio Pretraining on NUMBERS95**

![Diagram showing the steps of audio pretraining](image)

**Audio Labeling**

1. AV Database Audio
2. ANN Pretrained
3. Alignment
4. AV Database Audio Labeled

**Audio Training**

1. AV Database Audio Labeled
2. ANN Pretrained
3. ANN Audio Trained

**Audio Relabeling**

1. AV Database Audio
2. ANN Relabeled
3. AV Database Audio Relabeled

**Video Labeling**

1. AV Database Video
2. AV Database Video Labeled

**Fig. 3. Steps performed during the multi-stage labeling of the audio-visual database**

**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 (of Mons University, 1997)).

**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 by 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>train-</td>
<td>first</td>
<td>relabel-</td>
<td>video</td>
<td>only</td>
</tr>
<tr>
<td>11.6%</td>
<td>28.5%</td>
<td>7.1%</td>
<td>3.6%</td>
<td>33.1%</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, 7.1% and 28.5%, respectively). 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 10⁶ 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 recognition tests on AVNB95 at different SNR levels using the SI system as presented in Sec. 2 were performed to demonstrate the efficiency of the labeling process.

Finally we compared the two fusion methods, with (Eq. 3) and without (Eq. 6) the incorporation of priors.

In all cases training of the ANNs was performed on clean speech without additional noise.

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. Recognition tests with Gaussian white noise added at different SNR were performed. The results can be seen in Fig. 4. Over all SNR values an improvement due to the pretraining is visible. We performed a similar test where noise recorded in a car was added. In this case the pretraining lead to a deterioration of the recognition scores except for the clean speech case. Thus we concluded, that the amelioration of the recognition scores after pretraining is mostly due to the adaptation of the ANN to the noise. NUMBERS95 is recorded over telephone lines with low speech quality and hence significantly more noise is present than in our database. The characteristics of this noise are more similar to White noise than those of the noise recorded in a car. Consequently, when using the car noise for recognition the performance drops. Therefore we used for the following tests only the system directly trained on AVNB95.

4.2 Recognition Tests with a SI Structure

In order to evaluate the results of the labeling of the database and to demonstrate 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. For the tests, white Gaussian noise was added to the database. The fusion process was performed as detailed in Sec. 2.

4.3 Comparison of the Two Fusion Modalities

To evaluate the performance of our new audio-visual fusion modality using the a priori probability of the phoneme $H_i$ (see Eq. 3 and Eq. 4)
we compare it to the fusion without the use of a priori probabilities according to Eq. 6.

![Graph showing comparison of word error rates at different SNR's with and without a priori weighting.]

Fig. 6. Comparison of the word error rates at different SNR’s with white noise added for recognition with and without the use of a priori probabilities.

As can be seen in Fig. 6, our new fusion modality performs significantly better than the one without the use of a priori probabilities. The largest differences occur for medium SNR (8.5% at 3dB), as here both, the video and the audio path, have a significant contribution to the final a posteriori probability. When either the audio or the video path is dominant our formula approaches the one without the use of a priori probabilities and hence the performance is identical.

5. CONCLUSION

We presented a new methodology to label an audio-visual database and to set up an audiovisual 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.

A comparison between the fusion procedure presented by us, which includes a priori probabilities, and the one commonly employed in the literature was carried out. The results showed, that the recognition results obtained by using the a priori probabilities are particularly better for the situations of most concerne, where both, the audio and the video path, significantly contribute to the final a posteriori probability.

We further investigated the effects of integrating information from pretraining on a large audio database into the final system. The results depended on the type of noise used during testing. Therefore, in the case where the type of noise is known in advance, similar effects can be achieved via adding the noise already during the training process. Nevertheless pretraining is 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 and Christophe Savariaux 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 and Christian Sørensen for carrying out most of the simulations.

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

7. REFERENCES


