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
Universitt Karlsruhe
Kaiserstraße 12, 76128 Karlsruhe, Germany
{heckmann, kroschel}@int.uni-karlsruhe.de

[⊗] Institute de la Communication Parlée (ICP)
Institute 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]-[3]. This led to different approaches to integrate the visual information also in automatic speech recognition systems [4]-[10]. 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 [11]. This is all the more encouraging as they are up to now by far less elaborated as their Gaussian Mixture model based counterparts. 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 [12]. 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 [9] for more details).

Audio feature extraction is performed using RASTA-PLP [13] 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 move-

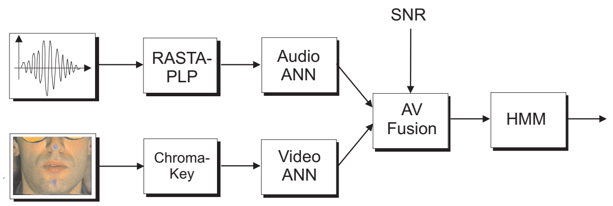


Figure 1: Separate Identification (SI) audio-visual speech recognition system

ment parameters can be extracted in real time. As lips parameters

- outer lip width
- inner lip width
- outer lip height
- inner lip height
- lip surface and
- mouth surface surrounded by lips

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. Hence a continuous recognition of numbers is performed. 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.

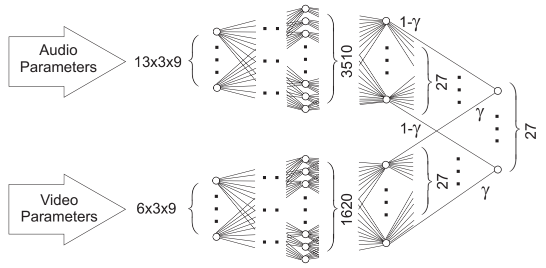


Figure 2: ANN used in the DI structure

In the SI structure during the audio-visual fusion process the final a posteriori probability $P(H_i|\mathbf{x}_A, \mathbf{x}_V)$ for the occurrence of the phoneme H_i is calculated from the two a posteriori probabilities $P(H_i|\mathbf{x}_A)$ and $P(H_i|\mathbf{x}_V)$ of the phoneme H_i estimated by the corresponding ANN's from the acoustic feature vector $\mathbf{x}_A$ and the video feature vector $\mathbf{x}_V$. To allow an adaptation of the fusion to the SNR we introduced the weighting factor γ, which is adjusted manually depending on the noise level [10]. The final a posteriori probability is thus calculated according to:

$$P(H_i|\mathbf{x}_A, \mathbf{x}_V) = \frac{P_A^{1-\gamma}(H_i|\mathbf{x}_A) P_V^{\gamma}(H_i|\mathbf{x}_V)}{\sum_{j=0}^{N-1} P_A(H_j|\mathbf{x}_A) P_V(H_j|\mathbf{x}_V)} \quad (1)$$

It is assumed that the SNR is known and hence no SNR estimation was implemented.

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 already experiences were 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 transfered. 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.

3.2 Multi-Stage Training

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. Common ways to achieve the labeling are a complete manual labeling, manual labeling of a subset of the database, training of the ANN on the subset and than using the ANN to label the rest of the database or 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 (compare 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

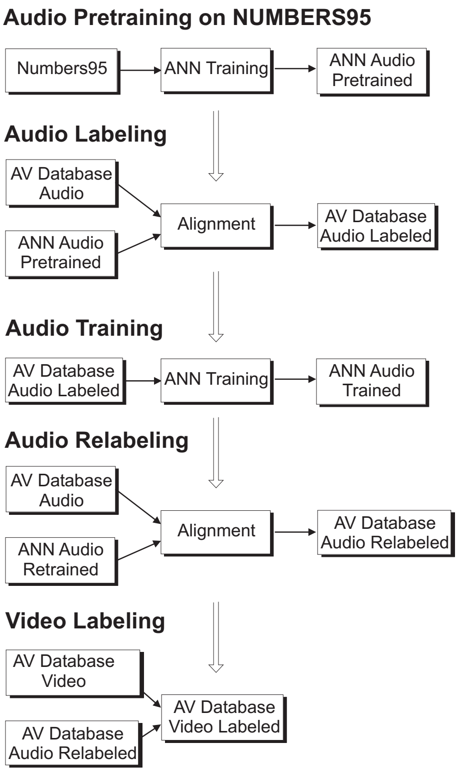


Figure 3: Steps performed during the multi-stage labeling of the audio-visual database

under varying conditions from the NUMBERS95 database and performed a cross-check during training on an independent test set to avoid over-adaption (compare [12]).

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 to 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, corresponding to the ANN, gives the highest overall likelihood is chosen.

	NB95	AVNB95 training on NB95	AVNB95 first labeling	AVNB95 relabeling	AVNB95 video only
Word Error Rate	11.6%	28.5%	7.1%	4.0%	35.1%

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 and do not occur synchronously with 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.

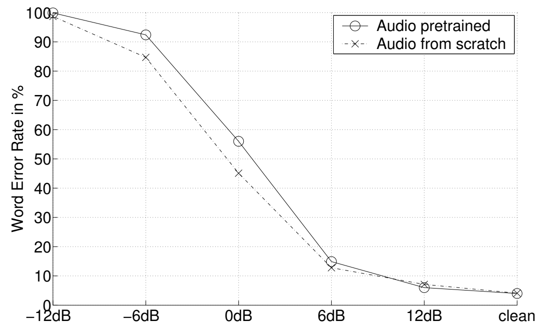
4 TRAINING AND TEST

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 weigths 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 one we trained directly on AVNB95. The results can be seen in Fig. 4. At high SNR's an improvement due to the pretraining of $1 - 2\%$ 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. The results of the pretraining for a SI structure are directly correlated to those of the audio part alone.



(a)

Figure 4: Comparison of the word error rates with and without pretraining at different SNR's for recognition with the audio part only

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.

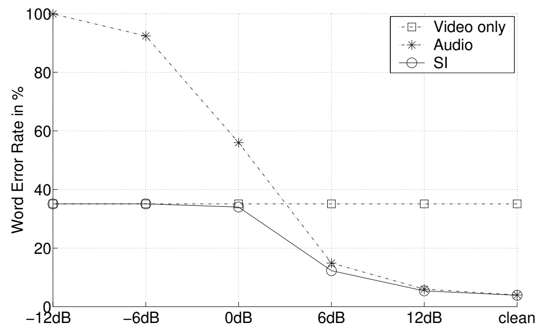


Figure 5: Comparison of the word error rates at different SNR's for recognition on the audio and video part only and with a SI structure

Fig. 5 compares the results of the audio and video only recognition with the audio-visual recognition. It can be seen that the audio-visual score is always better than the audio or video score alone. The synergy effect resulting from the audio-visual recognition is rather moderate, which is due to the simplified viseme model and the basic fusion strategy. These tests intent only to demonstrate the success of the labeling process and the good suitability of an hybrid ANN/HMM system for continuous audio-visual recognition, though. When looking at the word error rates it further has to be taken into account that a continuous word recognition was performed. Hence many errors are due to deletions and insertions which do not occur in a single word recognition.

5 CONCLUSION

To setup an audio-visual speech recognition system based on an ANN/HMM mixture approach, large training sets are necessary due to the huge size of the networks involved. The basic element of such a system is therefore an audio-visual database of considerable size. As labeled audio-visual databases having a sufficient size are not easily available and the desired audio and video recording conditions might be very different depending on the desired application, such a database often has to be recorded from scratch.

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. This made it possible to use the same phoneme models and the text transcription. 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, e.g. a more complex viseme model and a more elaborated fusion strategy, are necessary.

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 NUMBERS95, which was recorded over telephone lines, and our database. 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 speaker. Furthermore we want to thank Thorsten Wild, Vincent Charbonneau and Christian Sörensen for carrying out most of the simulations.

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

REFERENCES

[1] C. Benoit, T. Mohamadi, and S. D. Kandel, "Effects of phonetic context on audio-visual intelleginility of french", *J. Speech Hearing Res.*, vol. 37, pp. 1195–1203, 1994.

[2] N. P. Erber, "Auditory-visual perception of speech", *J. Speech Hear. Disord.*, vol. 40, pp. 481–492, 1975.

[3] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility", *J. Acoust. Soc. Amer.*, vol. 26, pp. 221–215, 1954.

[4] E. D. Petajan, *Adaptive Determination of Audio and Visual Weights for Automatic Speech Recognition*, PhD thesis, Univ. Illinois, Urbana, 1984.

[5] U. Meier, W. Hürst, and P. Duchnowski, "Adaptive bimodal sensor fusion for automatic speechreading", in *Proc. ICASSP'96*, Atlanta, GA, 1996, pp. 833–836.

[6] P. L. Silbsee and Q. Su, "Computer lipreading for improved accuracy in automatic speech recognition", *IEEE Trans. Speech Audio Proccessing*, vol. 4, pp. 337–351, 1996.

[7] S. Nakamura, R. Nagai, and K. Shikano, "Adaptive dtermination of audio and visual weights for automatic speech recognition", in *Proc. Eurospeech 97*, Rhodes, Greece, 1997, pp. 1623–1626.

[8] J. Luettin and S. Dupont, "Continuous auido-visual speech recognition", in *Proc. 5th Eur. Conf. Computer Vision*, 1998.

[9] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic spech recognition", *Speech Communication*, vol. 26, pp. 149–161, 1998.

[10] P. Teissier, J. Robert-Ribes, J.-L. Schwartz, and A. Guérin-Dugué, "Comparing models for audiovisual fusion in a noisy-vowel recognition task", *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 629–642, 1999.

[11] N.Morgan and H. Boulard, "Continuous speech recognition", *IEEE Sig. Proc. Magazine*, p. 24, May 1995.

[12] University of Mons, Mons, *Step by Step Guide to using the Speech Training and Recognition Unified Tool (STRUT)*, May 1997.

[13] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Rasta-plp speech analysis technique", in *Proc. ICASSP 1992 San Fransisco*, 1992, vol. 1, pp. 121–124.