SPHEAR Workshop 2000

# Labeling an Audio–Visual Database and Training an ANN/HMM Audio–Visual Speech Recognition system

SP EAR

Institut de la
Communication parlée
Frédéric Berthommier
Christophe Savariaux

Universität
Karlsruhe
Martin Heckmann
Kristian Kroschel

# Overview

- **The Database**
- **The System**
- **Multi–Stage Labeling**
- **Results**
- **Outlook**

Institut de la Communication
Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

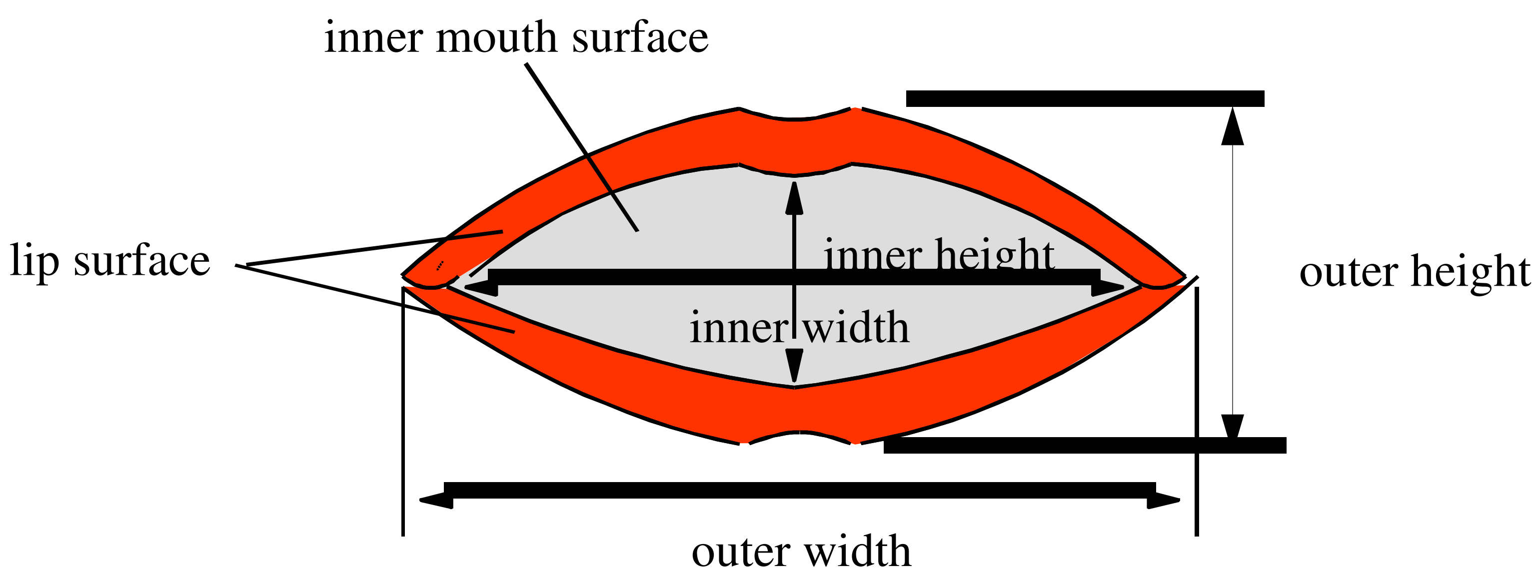
# Database Acquisition

- Transposition of NUMBERS95
  - **Audio–Visual repetition of a subset of NB95 (1700 sentences)**
- Audio and video recordings
  - **Chroma key process to extract lips parameters**

Institut de la Communication Parlée (ICP), Grenoble
F. Berthommier
C. Savariaux

Universität Karlsruhe
M. Heckmann
K. Kroschel

# Lip parameters





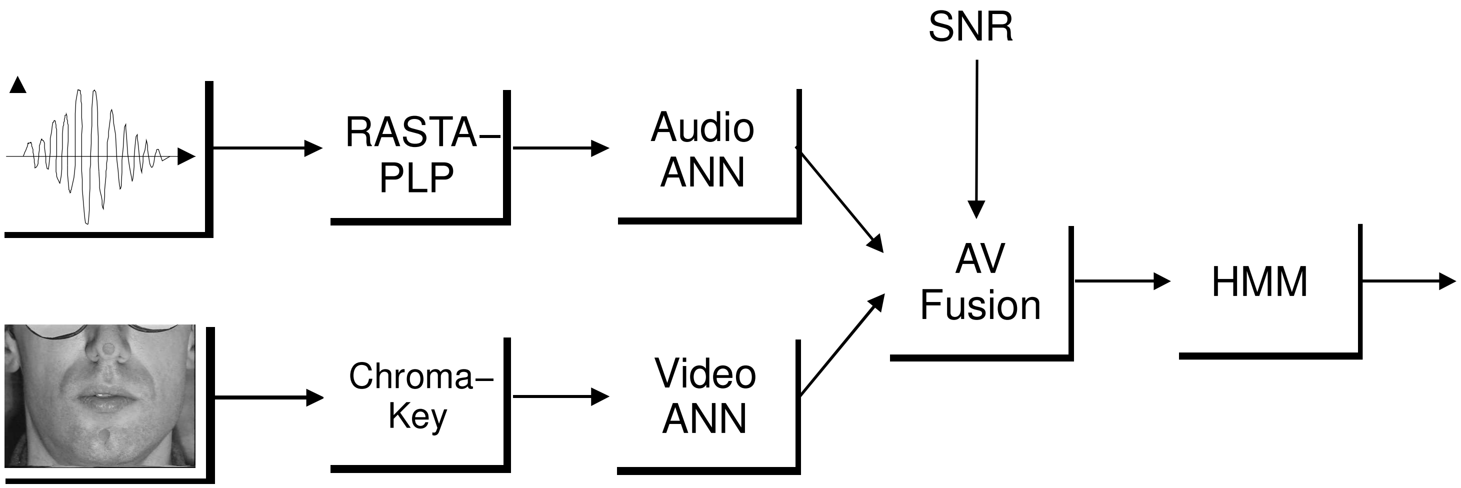
Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Recognition System




- Hybrid ANN/HMM audio–visual system (STRUT)
- Separate Identification (SI) structure
- SNR given as contextual information

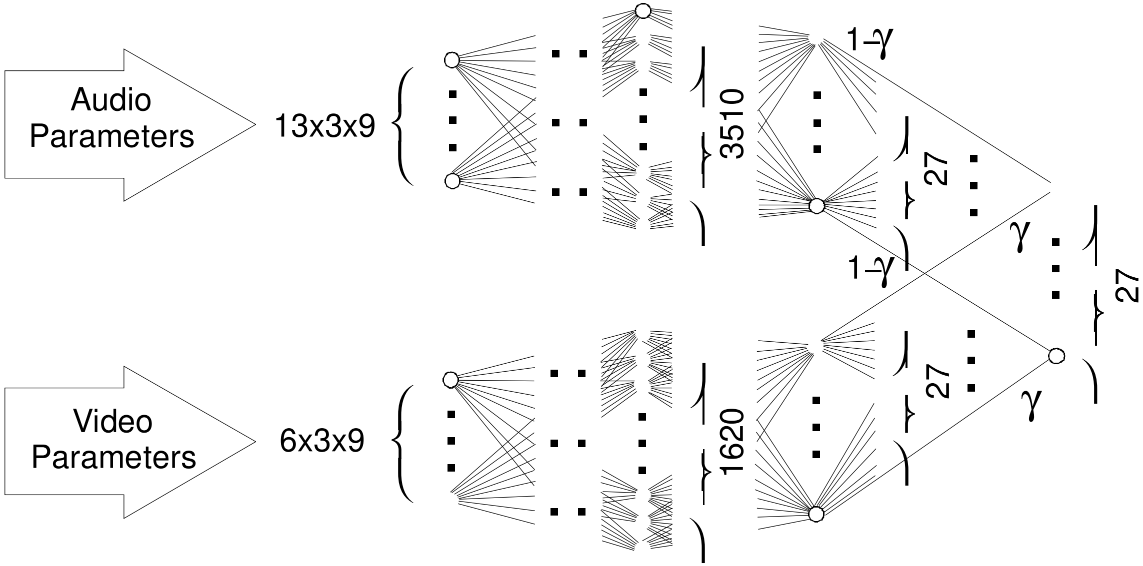

Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Recognition System




- Independent weighting of each posterior value for each frame is possible with STRUT

Institut de la Communication
Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

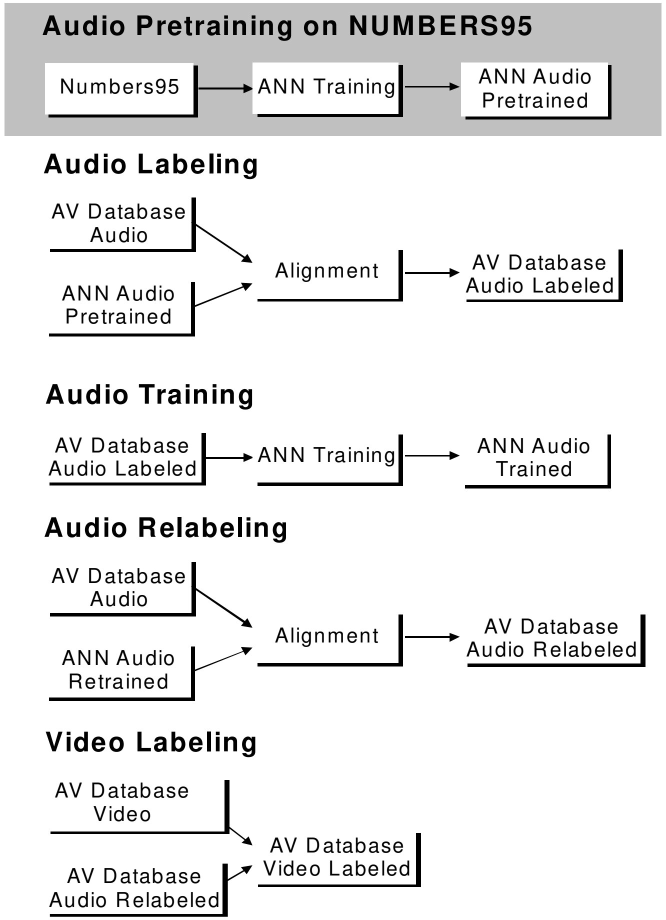
# Recognition System

- Fusion of audio and video via:

$$P(H_i \mid \mathbf{x}_A, \mathbf{x}_V) = P_A^{1-\gamma}(H_i \mid \mathbf{x}_A) P_V^{\gamma}(H_i \mid \mathbf{x}_V)$$

- Phoneme duration modeled via concatenation of states in the HMM
- Dictionary containing 30 words is used
- Grammar free continuous numbers recognition
- No distinction between phonemes and visemes is made

# Multi–Stage Labeling




- **Training on large multi speaker database NUMBERS95**
- **WER on NB95: 11.6%**

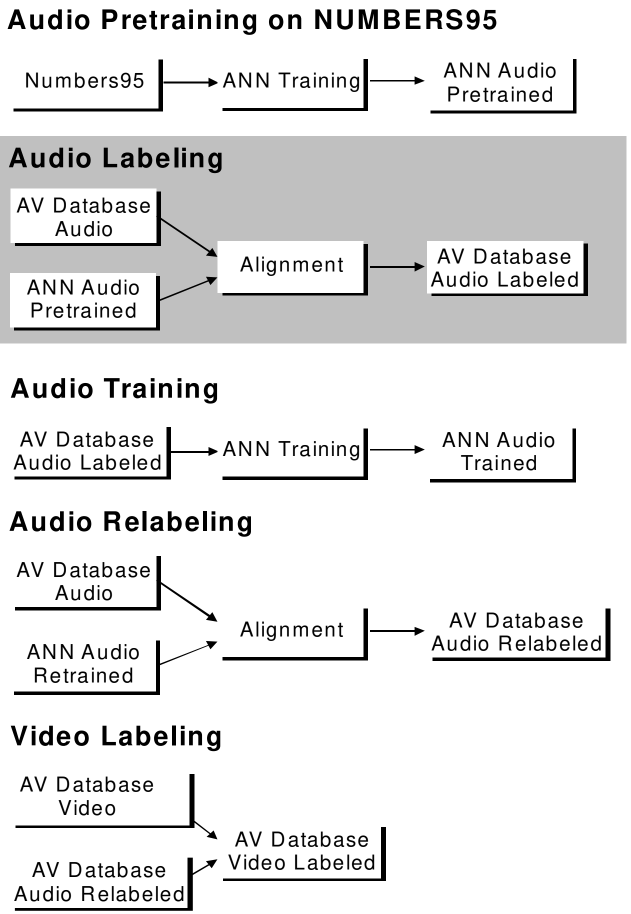

Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

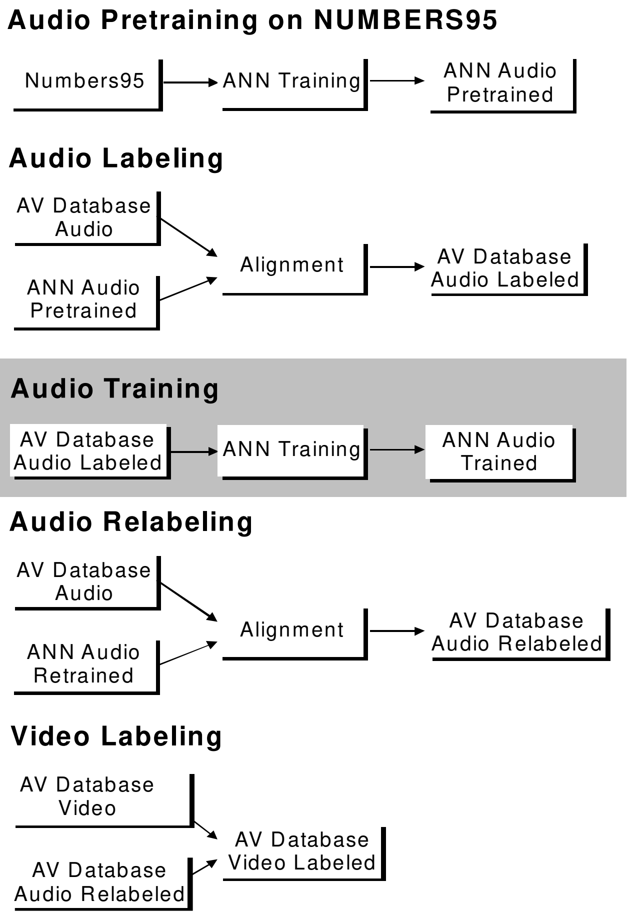
M. Heckmann
K. Kroschel

# Multi–Stage Labeling




- **Multi speaker NUMBERS95**
- **WER on NB95: 11.6%**
- **Single speaker Audio–visual NB95**
- **WER on AVNB95: 28.5%**
- **Forced Viterbi alignment**

# Multi–Stage Labeling




- **WER AVNB95 first labeling: 7.1%**

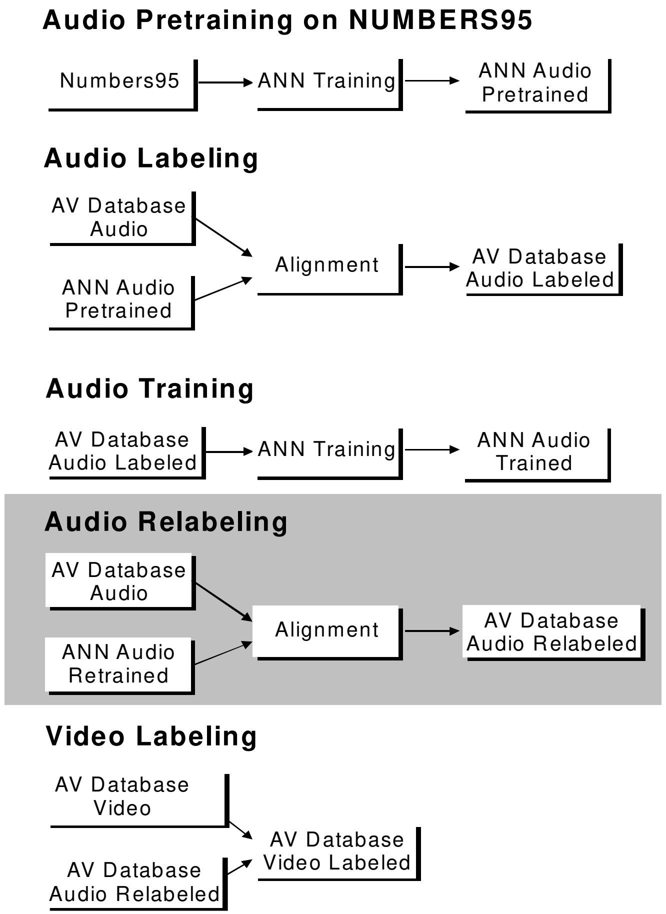

Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Multi–Stage Labeling




- **WER AVNB95 first labeling: 7.1%**
- **WER AVNB95 second labeling: 4%**

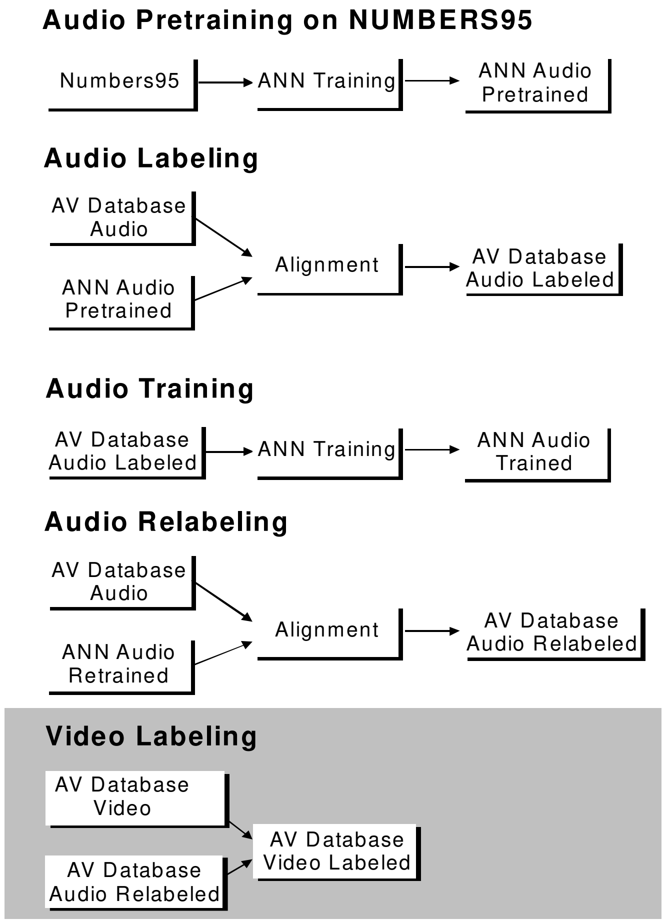
Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Multi–Stage Labeling




- **Direct application of the audio labeling to the video path**
- **WER using only video: 35.1%**

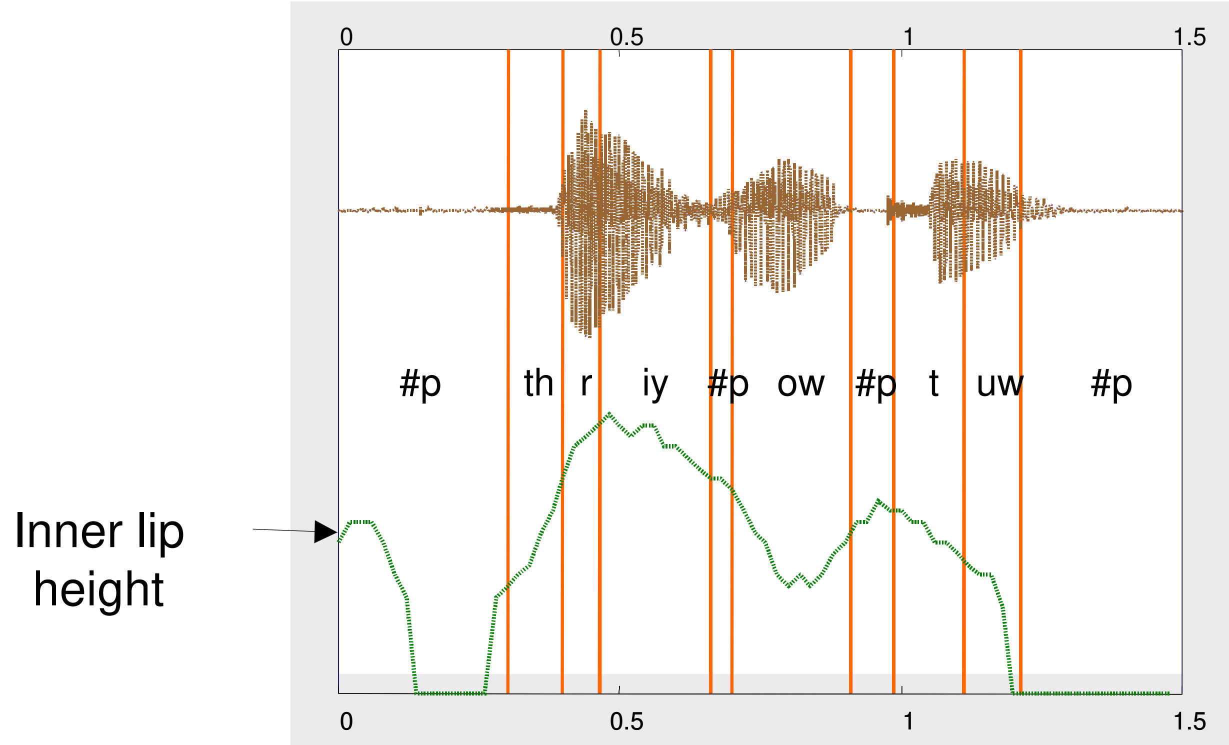
Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Result of the labeling




- Result of audio and video labeling for „three oh two“

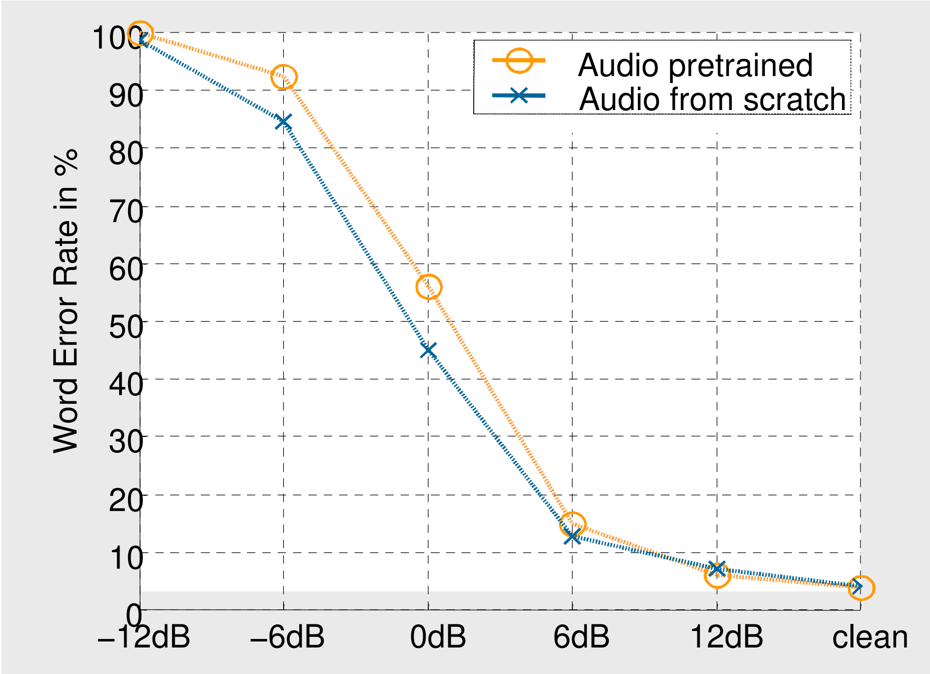
Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Effects of Audio Pretraining




- **Comparison between training from scratch on AVNB95 and pretraining on NB95 followed by training continuation on AVNB95**

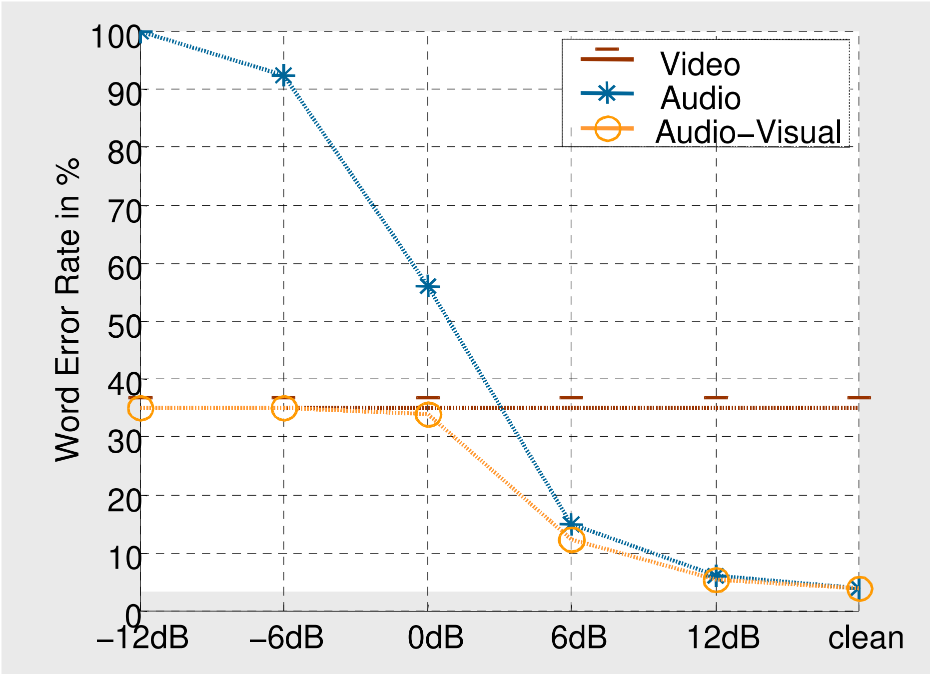
Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Audio–Visual Recognition




- **Comparison of audio, video and audio–visual recognition**

Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Conclusion

- Automatic audio–visual labeling process presented
- Pretraining on a large audio database only advantageous for labeling
- Good audio–visual recognition scores with hybrid ANN/HMM system


Institut de la Communication Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel

# Outlook

- Improvements of the recognition scores via
  - **Grouping of visually identical phonemes to visemes**
  - **Introduction of a more elaborated fusion modality**
- Implementation of a system to determine the weighting of audio and video from the data (e.g. entropy of posteriors) instead of using SNR given as context information

Institut de la Communication
Parlée (ICP), Grenoble

F. Berthommier
C. Savariaux

Universität Karlsruhe

M. Heckmann
K. Kroschel