Speech activity detection and speaker identification for CHIL

X. Zhu, C.C. Leung, C. Barras, L. Lamel, and J-L. Gauvain

LIMSI-CNRS, BP 133
F-91403 Orsay Cedex, France

1 Introduction

This paper presents some experimental studies at LIMSI on Speech Activity Detection (SAD) and Speaker Identification (SID) in the framework of the project, "Computers In the Human Interaction Loop" (CHIL). The objective of CHIL is to create environments in which computers serve humans who focus on interacting with other humans as opposed to having to attend to and being preoccupied with the machines themselves. For this project, SAD and SID technologies will be useful for a range of CHIL services. Experiments are reported on CHIL seminar data using either the close talking microphone channel (CTM) or one channel of the microphone array (ARR).

2 Speech activity detection

SAD is the process of dividing an input audio stream into speech/non-speech segments, and it is very useful as front-end for many audio technologies such as automatic speech recognition, speaker identification and verification, speaker localization etc. SAD is performed using two Gaussian mixture models (GMMs) respectively for speech and non-speech. A Viterbi decoder then provides the segmentation for the speech/non-speech labeling. The balance between Speech Detection Error Rate (SDER) and Non-speech Detection Error Rate (NDER) is reached using specific transition penalty between models.

For CHIL dry-run evaluation, the standard LIMSI SAD system for Broadcast News data was used. This system has a low SDER of 3.3% on close-talking microphone signal but a high NDER (cf. Table 1). For CHIL #1 evaluation, new GMMs were trained on available meeting data (ICSI, ISL, NIST). The Run #1 system has a 35% relative reduction of Average Detection Error Rate (ADER) for the Close-Talking data compared to the dry-run system. However, no improvement is observed on Far-Field data between two systems; better matched training data are needed to improve performance for the far-field condition.

3 Acoustic speaker recognition

The purpose of speaker recognition in the CHIL project is to recognize the identity of speakers, mainly the presenters of seminars. For the experiments,

Table 1. Comparison of Dry-run and Run #1 LIMSI SAD results (in %)

System	Channel condition	SDER	NDER	ADER
Dry-run	CTM	3.3	35.3	19.3
Eval #1	CTM	8.4	16.0	12.2
Dry-run	ARR	24.8	4.9	14.9
Eval #1	ARR	14.6	15.5	15.1

15 seminars were available, and the presenters of the seminars were the target speakers for the speaker identification task.

The speaker recognition system is a standard GMM-based system. A gender-independent universal background model (UBM) with 2048 Gaussian mixtures was trained using 7 hours of data from the ICSI, ISL, NIST meeting and the TED speeches corpora. Each target speaker model was trained by maximum a posteriori (MAP) adaptation of the Gaussian means of the UBM.

Training and testing utterances with different durations and channel conditions were used. Experimental results in speaker identification are given in Table 2. An obvious degradation of the identification performance due to the channel mismatch is observed. As expected, the performance is generally better when the speech duration is increased, and the close-talking microphone performed better than the microphone array.

Table 2. Identification errors for LIMSI SID system (in %)

	Test duration (sec)		60	30	10	5	1
Train duration (sec)	Train data	Test data					
60	CTM	CTM	0.0	0.0	0.1	1.7	17.6
60	CTM	ARR	14.5	17.3	37.2	47.1	70.9
60	ARR	CTM	20.9	19.0	15.9	23.0	55.5
60	ARR	ARR	9.1	8.2	12.9	17.2	44.0
30	CTM	CTM	0.0	0.0	0.1	2.0	19.2
30	CTM	ARR	16.4	20.9	33.7	43.0	67.7
30	ARR	CTM	13.6	14.3	18.6	28.7	63.7
30	ARR	ARR	1.8	0.0	4.7	11.3	52.8

4 Conclusions

The work at LIMSI related to speech activity detection and speaker recognition in the CHIL project has been briefly summarized. As expected, channel mismatch between close-talking microphone and far-field array microphone data is a major issue which needs to be addressed in future studies. However, both technologies provide already usable components for the CHIL project.