GOAL-DIRECTED ASR IN A MULTIMEDIA INDEXING AND SEARCHING ENVIRONMENT (MUMIS)

Mirjam Wester, Judith M. Kessens & Helmer Strik
A²RT, University of Nijmegen, The Netherlands

Introduction

The MUMIS project aims at creating a simple user-interface which can be used to query a video archive containing football matches.
For more info see:
http://parlevink.cs.utwente.nl/projects/mumis.html

This study concentrates on the automatic speech recognition (ASR) portion of the project. The data comprises commentaries accompanying the EURO-2000 football matches for:

- Dutch
- English
- German

The goal of this study is to illustrate the feasibility of using ASR technology to transcribe the spoken commentaries that accompany football broadcasts.

Speech Material

The commentator(s)'s speech was orthographically transcribed by SPEX. Test material was taken from the Yugoslavia - The Netherlands match.

Table 1: *Number of words in test set, per language.*

	Dutch	English	German
¼ of match	1577	2641	1000

Table 2: *Sets of training material available per language.*

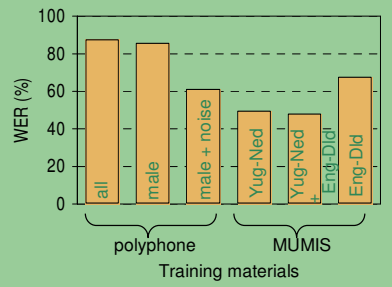
language	material	selection	duration speech	Mean SNR dB (std)
Dutch	Polyphone	all speakers	24h 37m	36.1 (5.3)
		male speakers	12h 32m	36.6 (5.4)
	MUMIS	Yug-Ned	19m	9.4 (2.9)
		Eng-Dld	18m	8.2 (2.8)
English	MUMIS	Yug-Ned	29m	12.1 (3.6)
		Eng-Dld	34m	10.8 (3.3)
German	Mumis	Yug-Ned	14m	10.9 (2.7)
		Eng-Dld	24m	6.9 (1.7)
		Dld-Rom	21m	7.4 (2.5)
		Port-Dld	33m	8.5 (2.9)
		Ned-It	24m	9.2 (2.8)
		Port-Tur	30m	19.5 (4.2)
		Fra-It	34m	11.3 (3.4)
		Eng-Port	23m	15.2 (4.3)

Speech Recognizer

Standard HMM system; features are 14 MFCCs + deltas
Acoustic Models: sets of phones derived from SAMPA: English 40, Dutch 37 and German 34, plus a non-speech model.
Lexicon: Most transcriptions derived from CELEX, missing words were transcribed manually.
Language Model: Trained on orthographic transcriptions for each match.

Dutch results

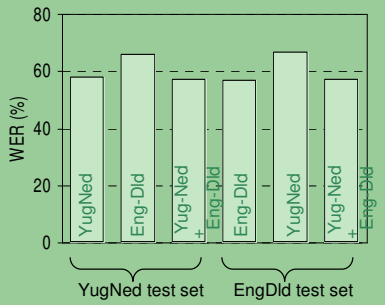
Results for different sets of training materials.



- Matched training/test data leads to the best performance.
- Using Polyphone training data does not lead to lower WERs than using MUMIS training data; even though it comprises more data.
- Noisifying Polyphone data with MUMIS noise leads to a substantial improvement.

English results

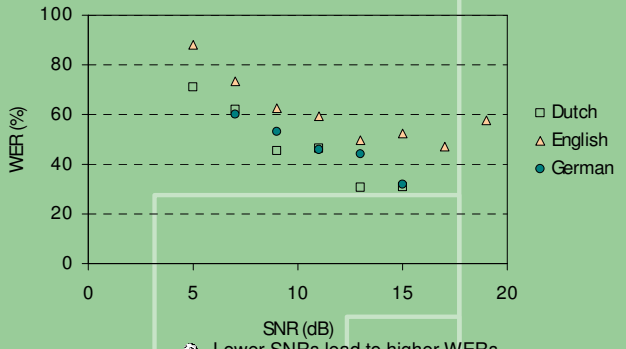
Results for cross-testing Yug-Ned and Eng-Dld.



- Matched training/test data leads to best performance.
- Cross-testing shows increase in WERs.

WER for different SNR values

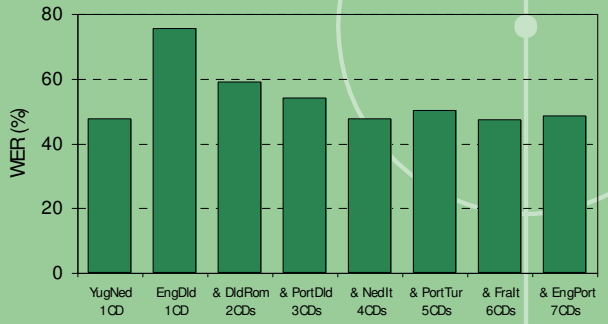
For each utterance the SNR was calculated. The utterances were grouped into 2dB wide bins according to their SNR values, and the WER for each bin was calculated. Only bins containing > 100 words were considered.



- Lower SNRs lead to higher WERs

German results

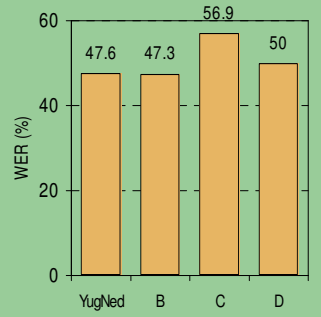
Results of adding additional MUMIS data.



- At first adding more data leads to lower WERs, however a floor seems to be reached after adding about 90 minutes training data.
- The training data was added as it became available... However, a selection based on matching or similar SNR values may be better.
- See Table 3 for the various selections.

Table 3: *Selection of matches according to average SNR values.*

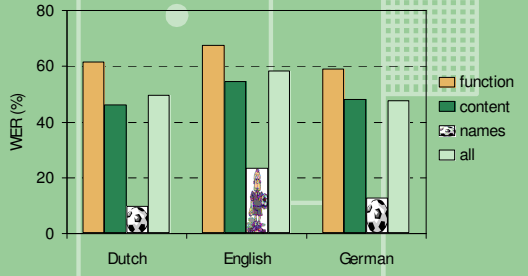
SNR classification	matches
A=very noisy	EngDld+DldRom
B=noisy	PortDld, NedIt + FraIt
C=semi-clean	PortTur + EngPort
D=minus very noisy	noisy and semi-clean



- The best result is obtained when the training data matches the test data, i.e. YugNed, or when matches are selected for training with similar SNR values, i.e. the noisy selection (B).

WER for different word types

Split words into categories, i.e. football player's names, content words and function words.



- Function words cause a great deal of the errors (and make up 50% of the words).
- The application specific words, players' names, are recognized best.

Conclusions

- SNR values explain the WERs to a large extent.
- More data is not necessarily better.
- Matching the SNR values of training and test material leads to best results.
- Overall WERs are very high, but application specific words are recognized quite well.

Ongoing work

- Generic language model and lexicon.
- Speaker adaptation.
- Transcription of more German data.

Acknowledgments: This work was carried out in the context of MUMIS, an EU funded project within the Information Society Program of the European Union, section Human Language Technology. (IST-10651)