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

Mirjam Wester, Judith M. Kessens & Helmer Strik

A²RT, Department of Language and Speech
University of Nijmegen, The Netherlands
{M.Wester,J.Kessens,Strik}@let.kun.nl

ABSTRACT

This paper describes the contribution of automatic speech recognition (ASR) within the framework of MUMIS (Multimedia Indexing and Searching Environment). The domain is football commentaries. The initial results of carrying out ASR on Dutch and English football commentaries are presented. We found that overall word error rates are high, but application specific words are recognized reasonably well. The difficulty of the ASR task is greatly increased by the high levels of noise present in the material.

1. INTRODUCTION

This paper reports on the automatic speech recognition research that has been carried out in the context of the Multimedia Indexing and Searching Environment (MUMIS) Project (IST 10651) [11]. MUMIS is an ongoing EU-funded project within the Information Society Program (IST) of the European Union, section Human Language Technology. MUMIS is concerned with the development and integration of base technologies, demonstrated within a laboratory prototype, to support automated multimedia indexing and to facilitate search and retrieval from multimedia databases. It makes use of data from different media sources (textual documents, radio and television broadcasts) in different languages (Dutch, English, and German) within the domain of football (UEFA tournament 2000, referred to as EURO-2000). The type of access to the video material that is envisaged by the MUMIS project is that with the help of a simple user interface formal queries can be input to the system such as "show me all the goals by Beckham" or "show me all the Dutch penalties". One of the essential technologies that is employed in MUMIS is automatic speech recognition (ASR). The ASR component is vital to time-aligning information gained from textual sources and spoken sources to the relevant images in the video data. The purpose of this paper is to report on ongoing work concerning the ASR portion of the project. For further information on the MUMIS project see [3, 4, 11].

To date, recognition experiments have been carried out for English and Dutch. The data that has been used comprises the commentaries that accompany TV broadcasts of the EURO-2000 football matches. The speech can be described as spontaneous. The recordings are extremely noisy as they contain a great deal of background noise produced by the crowd, the referees etc. This noise greatly increases the difficulty of the speech recognition task. On the other hand, the task is maybe less difficult than some other ASR tasks as each match only has one or two speakers.

In Section 2, the type of speech material that we have been working with is described. Next, a short summary of the ASR system is given. This is followed by the results obtained in recognition experiments and an analysis of these results. We conclude by discussing the findings and by explaining the further steps that will be taken to produce improved word transcriptions of spoken football commentaries.

2. SPEECH MATERIAL

The speech material, used in the current experiments, was recorded from TV broadcasts of the football matches at the Max Planck Institute for Psycholinguistics, Nijmegen. Two matches for both British English (ENG) and Dutch (NL) were orthographically transcribed by SPEX [12]: Yugoslavia – The Netherlands (Yug-Ned) and England-Germany (Eng-Dld). In the recordings, stadium noise and the commentator's speech are mixed. Note that only the speech of the commentator(s) has been transcribed.

2.1. Transcription

The transcriptions delivered by SPEX contain the orthographic transcription of the commentator's speech and an alignment at chunk-level. A chunk is a segment of speech of about 2 to 3 seconds. The CGN (Corpus Gesproken Nederlands) [13] orthographic transcription protocol was adhered to during the transcription and chunk-alignment of the MUMIS data. Relevant points from the CGN protocol in the context of MUMIS are:

- The orthographic transcription is to be an exact account of what a speaker has said at the word level.
- The transcription is to be closely linked to standard written language, i.e. various ways of pronouncing words is not reflected in the spelling.
- Separate tiers indicate different speakers.
- A system of codes is used to indicate irregularities in the speech. For instance, relevant transcription codes that were encountered in the MUMIS transcriptions are: *v for foreign words, *a for truncated words, *x for words which were not properly understood, and *u for disfluencies.

2.2. Statistics of the MUMIS material

Table 1 and 2 show statistics that were calculated on the basis of the transcribed matches. The tables show the total number of chunks, the number of chunks containing speech, and the number of chunks containing non-speech. The non-speech chunks are not empty but contain stadium noise. For the English

data, given in Table 2, the speech chunks are listed for two speakers (sp 1 and sp 2) because the commentary is given by two speakers. The English data also contains a number of chunks of overlapping speech. Finally, the number of words (types and tokens) and the average number of words per chunk are also given.

Comparing Table 1 and Table 2 illustrates a number of differences between the commentaries of the two languages. First of all, there are two speakers in the English commentaries and only one in the Dutch material. Listening to the material, we observed that the English commentary is more dialogue-like, whereas the Dutch speech material is clearly a monologue. Furthermore, the English material contains more speech than the Dutch material which can be deduced from the number of word tokens. However, not only is there more speech, the English commentary also contains 1/3 more different words than the Dutch commentaries.

	Yug-Ned	Eng-Dld	Total
total # chunks	2511	2635	5146
# speech chunks	1473	1533	3006
# non-speech chunks	1038	1102	2140
# words (types)	1304	1174	1954
# words (tokens)	6123	5956	12079
average # words per chunk	4.2	3.9	4.0

Table 1: Statistics of the Dutch material.

	Yug-Ned	Eng-Dld	Total
total # chunks	2564	3049	5613
# speech chunks	1724 (total) 1144 (sp 1) 580 (sp 2)	2001(total) 1398 (sp 1) 603 (sp 2)	3725
# non-speech chunks	821	1022	1843
# overlapping chunks	19	26	45
# words (types)	1769	1985	2923
# words (tokens)	10335	13687	24022
average # words per chunk	6.0	6.8	6.5

Table 2: Statistics of the English material.

2.3. Training and test material

For training purposes, data was taken from the above described material and from the Dutch Polyphone database [2], a 5000 speaker telephone database. A random selection of each MUMIS match was made dividing the match into ¾ for training and ¼ for testing the system. The number of words and chunks in the test sets are shown in Table 3. From the Polyphone database, we selected the phonetically rich sentences read by male speakers since the commentators in the matches are all male speakers.

Test material	#words	#chunks
Yug-Ned_NL	1577	626
Eng-Dld_NL	1401	653
Yug-Ned_ENG	2641	456
Eng-Dld_ENG	3311	500

Table 3: Number of words and chunks in the various test sets.

Table 4 shows the different sets of training material and the corresponding duration of speech material. Estimations of the durations of the speech in the training material are based on automatic segmentations and the non-speech frames were disregarded in the calculation of these durations. In Table 4 it can be seen that the amount of speech data present in one match is not very large. A match lasts about 90 minutes, and from the training material durations (3/4 of a match, 68 min) we can deduce that the Dutch commentator only speaks for about 30% of the time, and the English commentators speak for about 50% of the time. Consequently, the amount of speech material available to train MUMIS models is not comparable to the usual quantities used for training acoustic models. Therefore, also Polyphone data was used to train Dutch acoustic models.

	Material	Selection	Duration
Dutch	Polyphone	male speakers	12h:32 min
	MUMIS	Yug-Ned	19 min
	MUMIS	Eng-Dld	18 min
English	MUMIS	Yug-Ned	29 min
	MUMIS	Eng-Dld	34 min

Table 4: Overview of the selections of training material and the corresponding duration of speech present in the sets of materials.

2.4. Signal to Noise ratios (SNRs)

For an indication of the noise present in the MUMIS material, we calculated signal-to-noise ratios (SNR) for the chunks that contained speech. For the sake of comparison the SNR for Polyphone data was also calculated. The signal energy was calculated by taking the 70% frames with the highest root mean squared energy values, and the noise energy was calculated over the remaining 30% frames. The SNR in dB is 10 times the log of the signal energy divided by the noise energy.

Table 5 shows the mean SNR values and the standard deviation for each match per language. These SNR values clearly indicate that the matches are very noisy. English data has higher SNRs than Dutch data, indicating lower noise levels. Furthermore, it can be inferred from these values that the recordings of the Eng-Dld match are slightly noisier than the Yug-Ned match recordings.

Material	Mean	std
Yug-Ned_NL	9.4	2.9
Eng-Dld_NL	8.2	2.8
Yug-Ned_ENG	12.1	3.6
Eng-Dld_ENG	10.8	3.3
Polyphone	36.6	5.4

Table 5: SNR values in dB for the different sets of data.

3. CONTINUOUS SPEECH RECOGNIZER

The continuous speech recognition (CSR) system which was used for both languages is the Phicos system [10], a standard hidden Markov model (HMM) based system. Different sets of phone models, lexicons, and language models have been trained for the two languages, English and Dutch. The front-end acoustic processing consists of calculating 14 MFCCs plus their deltas, every 10 ms for 16 ms frames.

3.1. Acoustic models

For the English recognition system a set of 40 phones was used. The set was derived from the set of SAMPA symbols for British English [14]. For Dutch, a set of 37 phones was used [15]. In addition to the sets of phone models for both languages, a model for non-speech was trained. In effect, the non-speech model for MUMIS data is a noise model as all non-speech chunks contain noise. The acoustic models are continuous density monophone HMMs with 32 Gaussians per state. Each HMM consists of six states, three parts of two identical states, one of which can be skipped. The non-speech model consists of just one state [10].

3.2. Lexica

The Dutch phonetic transcriptions were automatically generated using the transcription module of the Text-to-Speech (TTS) system developed at the University of Nijmegen [4]. In this transcription module, phone transcriptions of words were obtained by looking up the transcriptions in two lexica: CELEX [1] and ONOMASTICA [9], a lexicon with proper names. For those words for which no transcription was available a grapheme-to-phoneme converter was used. All transcriptions were manually checked and corrected when necessary.

The English phonetic transcriptions were also mainly obtained from CELEX, roughly 65% of the words were directly taken from CELEX. The remaining 35% of the full list of words could be classified either as misspelled words (4%), names (7%), disfluencies (5%) or conjugations/inflections of words (19%). The misspelled words were corrected. The transcriptions for most conjugations (16% of the full list of words) were derived automatically from CELEX. For the names and the remaining words without a phonetic transcription (3%) the transcriptions were made by hand.

3.3. Language models

As a starting point, language models were trained on the basis of the chunk-level orthographic transcription for each of the matches. In the near future, a generic language model will be developed. The idea is that for each match it will be possible to adjust the language model by "slotting in", for example, the list of players, the name of the stadium and the countries involved in the match.

4. RESULTS

A number of recognition experiments were carried out using different sets of phone models to investigate how well this type of noisy football material could be recognized. In Table 6, the results for the Dutch commentaries are shown. The first column shows the test set in question, the second column gives the type of material that was used to train the acoustic models and the final column shows word error rates (WER= (substitutions + insertions + deletions) / total number of words).

These results show that a *mismatch* between the training and test data hurts recognition performance significantly. Training on Polyphone and testing on MUMIS leads to WERs in the 80s whereas when the training set and test set are matched the WER is around 50-60%. The high WERs for Polyphone can be explained by the large mismatch in SNR. The cross test conditions show a similar pattern, in that a mismatch leads to higher WERs. For the Yug-Ned match the WER goes up more than 10% when the acoustic models are trained on Eng-Dld. For the Eng-Dld test set the increase is less pronounced but there is still an increase. It is not clear why there is such a large difference in WERs between the result for Yug-Ned compared to Eng-Dld .

Test set	Training Material	WER(%)
Yug-Ned	Polyphone	85.4
	Yug-Ned	49.5
	Eng-Dld	60.2
Eng-Dld	Eng-Dld	62.9
	Yug-Ned	67.3

Table 6: *WER results for the Dutch data, for different sets of training material for both the Yug-Ned and Eng-Dld test sets.*

Table 7 shows the results obtained for the English commentaries. In this case, no Polyphone models have been trained, only MUMIS models. The same trend can be observed as for the Dutch material, that is, the WERs are high in the matched conditions but even higher in the mismatched conditions.

Test set	Training material	WER(%)
Yug-Ned	Yug-Ned	58.2
	Eng-Dld	65.9
Eng-Dld	Eng-Dld	57.0
	Yug-Ned	66.8

Table 7: WER results for the English data, for different sets of training data for both the Yug-Ned and Eng-Dld test sets.

When calculating the WERs all words are weighted equally. However, it is not the case that all words are equally important in an information retrieval task. In fact, it has been shown that disregarding commonly occurring function words in an information retrieval task improves the retrieval performance [8]. Furthermore, [7] reports that although WERs are high (around 55%) it is still possible to create a highly usable index. Additionally, speech recognition research has shown that function words make up a large portion of any spontaneous speech corpus and that they are notoriously variable in their pronunciation [5].

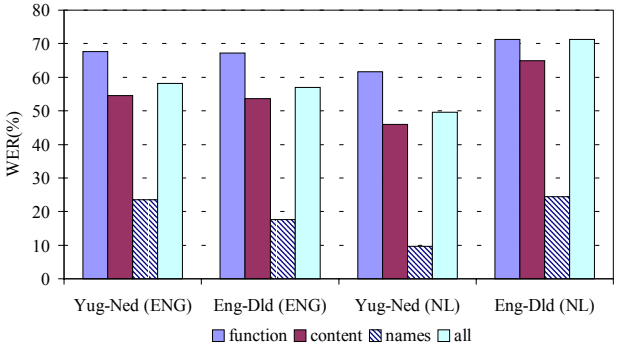

Figure 1: WERs according to word type.

Keeping these points in mind, we categorized the words in the lexicon into several word types. Three categories were defined: 1) function words, for instance, prepositions, pronouns, determiners etc., 2) an application specific category: "names of players", and 3) other content words. Next, the WER for each of

these groups of words was calculated. These results are shown in Figure 1. It can be seen that function words are badly recognized (incidentally, they make up roughly 50% of all words). The content words are recognized better than the function words which is encouraging, and the application specific words "players' names" are recognized quite well.

Furthermore, we wanted to ascertain whether the WERs are truly influenced by lower SNRs. To this end, we studied the relationship between the SNRs and the WERs for the MUMIS test sets in matched conditions. For each utterance, the SNR was calculated. The utterances for each match were sorted into bins according to their SNR values (each bin has a width of 2 dB), and the WER for each bin was calculated. Figure 2 shows WERs in relation to SNRs, only bins containing at least 100 examples have been plotted in the figure. The results are in line with what one would expect, i.e., lower SNRs lead to higher WERs. In addition, this figure shows a floor effect at roughly 30% for Dutch and at about 40% for English.

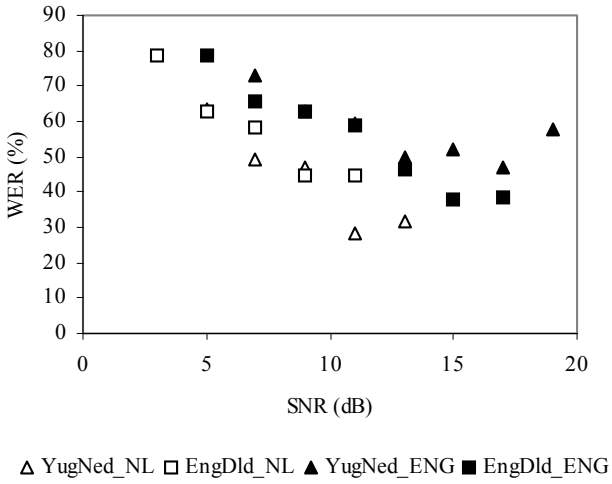

Figure 2: WERs for different SNR values.

5. DISCUSSION

The WERs that were obtained in our experiments are undesirably high. Consequently, lowering the WERs is one of the main objectives of continued research. However, information extraction (IE) experiments using this impaired ASR output will determine to what extent the present WERs may be acceptable. As Figure 1 showed, the WERs for application specific words are much lower than the overall WERs. For IE, it is not evident that function words are needed to obtain the correct fragments in the video. Merging the different sources (text and annotations from commentary) will show whether or not IE suffers from incorrect automatic transcriptions.

In improving the WERs, one of the issues we will be dealing with is the lack of appropriate training material. Presently, there is not enough MUMIS material to properly train acoustic models. Therefore, more of the EURO-2000 games will be transcribed by SPEX. In addition to more data for Dutch and English, German commentaries will also be transcribed. The other solution to the lack of material is to exploit Polyphone differently. At present the Polyphone data has not been very useful because the mismatch between the two sets of data is too large, not only in terms of SNR, but also in terms of speech style. Polyphone is read speech, whereas MUMIS is spontaneous speech. Possibly, using adaptation techniques in which Polyphone models are adapted with MUMIS material, will be more fruitful utilization of the Polyphone data.

Figure 2 showed us that there is a clear relationship between SNR values and WERs. Therefore the other main issue that we will be addressing in more detail is applying techniques for noise robust ASR to this data.

6. ACKNOWLEDGEMENTS

Eric Sanders (SPEX) is thanked for providing the tool that was used to calculate the SNR values.

7. REFERENCES

[1] Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2)* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

[2] Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. & Boves, L. (1994). Creation and analysis of the Dutch Polyphone corpus. *ICSLP '94*, Yokohama, Japan, pp. 1803-1806.

[3] de Jong, F. & Westerveld, T. (2001). MUMIS: multimedia indexing and searching. *Proc. of the Content-Based Multimedia Indexing Workshop* pp. 423-425.

[4] Declerck, T., Wittenburg, P. & Cunningham, H. (2001). The Automatic Generation of Formal Annotations in a Multimedia Indexing and Searching Environment. *Proc. of ACL/EACL Workshop on Human Language Technology and Knowledge Management* pp. 129-136.

[5] Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication, 29*, 159-176.

[6] Kerkhoff, J. & Rietveld, T. (1994). Prosody in NIROS with FONPARS and ALFEIOS. *Proc. of the Dept. of Language & Speech, University of Nijmegen*, Nijmegen, pp. 107-119.

[7] Logan, B., Moreno, P., Van Thong, J.-M. & Whittacker, E. (2000). An experimental study of an audio indexing system for the web. *ICSLP-00*, Beijing, China, pp. 676-679.

[8] Ng, K. (2000). Towards and integrated approach for spoken document retrieval. *ICSLP-00*, Beijing, China, pp. 672-675.

[9] Quazza, S. & van den Heuvel, H. (2000). The use of lexicons in test-to-speech-systems. In F. v. Eynde & D. Gibbon (Eds.), *Lexicon Development for Speech and Language Processing* (pp. 207-233): Kluwer Academic Publishers.

[10] Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C. & Geller, D. (1993). The Philips Research System for Large-Vocabulary Continuous-Speech Recognition. *Eurospeech '93*, Berlin, Germany, pp. 2125-2128.

[11] http://parlevink.cs.utwente.nl/projects/mumis.html

[12] http://www.spex.nl

[13] http://lands.let.kun.nl/cgn/ehome.htm

[14] http://www.phon.ucl.ac.uk/home/sampa/english.htm

[15] http://www.phon.ucl.ac.uk/home/sampa/dutch.htm