Proceedings of the 6th Nordic Signal Processing Symposium - NORSIG 2004
June 9 - 11, 2004
Espoo, Finland

Annotation and Automatic Recognition of Spontaneously Dictated Medical Records for Norwegian

Vidar Markhus, Bojana Gajic, Jacques Svarverud, Lars Erik Solbraa, and Magne H. Johnsen
Norwegian University of Science and Technology

NTNU
Department of Electronic and Telecommunication
O.S. Bragstads plass 2B,
N-7491 TRONDHEIM
NORWAY
E-mail: [vidarma, gajic, mhj]@iet.ntnu.no
and [svarveru, larserso]@stud.ntnu.no

ABSTRACT

In this paper we present a new research database of spontaneously dictated Norwegian speech, called MOBELspon, together with an experimental evaluation using standard automatic speech recognition (ASR) techniques. MOBELspon contains about 150 minutes of spontaneous dictation and 48 minutes of read speech of rheumatism health care records. The speakers are 10 medical students of both genders, coming from different parts of Norway and talking with their own dialect. MOBELspon contains a high degree of spontaneous speech features like disfluencies and para-linguistic speaker generated noise sounds. To model these features we propose some new special annotation symbols.

MOBELspon contains the highest degree of spontaneous speech features measured (11.6 % of transcribed sounds for the spontaneous speech part of the corpus are non-words) developed for Norwegian, making it a starting point for developing and testing more sophisticated techniques, and to model natural speech and dialects. Preliminary results show low recognition accuracy (as expected).

1. INTRODUCTION

ASR of spontaneous speech has been an active research area internationally in the recent years [1]. Although high recognition accuracy can be obtained using state-of-the-art ASR for speech in the form of reading a written text or similar, the accuracy is quite poor for freely spoken spontaneous speech. This is due to the fact that most of the earlier ASR systems where trained on speech read from written text, which has a different linguistic form than naturally spoken languages. Benchmark tests of large corpus of spontaneous speech, like Japanese [2], with acoustic and linguistic modeling of the spontaneous speaker generated sounds, are far more effective than models based on read speech, when the test corpus is natural language. Still, the recognition accuracy is to low for wide area commercial use of (highly complex) natural language speech recognizers.

The database is collected based on cooperation between the two projects BRAGE [3] and MOBEL [4]. The BRAGE project is the first large scale project that does research on spontaneous speech for the Norwegian language. The MOBEL project is a multidisciplinary research project that aims at specifying a context/procedure-aware, multi-modal (e.g. natural language) interface to the electronic patient record (EPR), that supports cooperative, patient-focused work. The MOBELspon database contains spontaneously dictated health care records for rheumatism. The BRAGE project needed a pilot project for testing and getting experience with more detailed annotating of spontaneous speech, and further, to evaluate the possibility of adapting existing acoustic models [5] to cope with spontaneous speech.

It has been invested enormous amount of money to get hospitals in Norway "paperless", i.e. to go from paper based patient records to fully integrated EPR systems. This paradigm shift is far from complete mainly because of the complexity of such systems, but also because health care workers prefer to work with paper records, even if the EPR has the required functionality [6]. This is caused by the lack of good usability in the systems. ASR used for dictation and as a part of a multi-modal dialog interface, may give increased usability and thereby make the EPR systems more appropriate to use. A presumption to make this feasible is that the ASR systems must cope with most of the dialects in Norway and speaker variability, since all workers have to use the EPR to eliminate the need for paper based patient records.

2. DATABASE DESCRIPTION

2.1 Database Collection

We used two different procedures for the recording. First, the medical students studied four or five patient case histories each. These cases were originally developed as practical problems used in a course in rheumatism at NTNU and contained only head words based on results from clinical examinations. Afterwards they dictated the whole patient record for each patient using only a minimum of notes without any pauses in the recording. Since the participants had to think of what they were going to say, formulate the sentences and dictate at the same time, they generated a lot of hesitations, disfluencies and other phenomena that are typical for the spontaneous speaking style. It should also be taken into account that medical students talked with their own dialect and were not familiar with dictation. Further, professional doctors usually take pauses in the recording after almost each sentence when they dictate. Thus, this can be seen as a "worst case" of what future dictation systems must handle.

In addition, each speaker was asked to read about 5 minutes from written rheumatism patient records. These records were prepared by a medical physician to make them anonymous. The participants read these patient records in bokmål (one of two written language for Norwegian), because they were written in this language.

The recording was done in an ordinary office, with a standard (low-cost) close-talking microphone. The speech was recorded in Microcoft Wave format, with sample frequency 16kHz. Some noise from bypassing cars and distant conversation can be heard on parts of the recordings.

2.2 Annotating the speech corpus

The text transcribed is annotated using a new standard for Norwegian spontaneous speech corpus, proposed by the BRAGE project [3]. We used the SpeechDat annotation standard [7] as a starting point. An important reason for this choice is that we have experience with it and we have a "Norwegian" interpretation of it. The SpeechDat annotation standard is not particularly accommodated for spontaneous speech, therefore we have proposed an expansion of it. Phonemes are transcribed using the SAMPA phonetic alphabet [8]. In addition, we are using special annotation symbols to distinguish between four different types of pronunciation error, four types of speaker generated noise and two types of environmental noise. No prosody information is used. A detailed explanation is given in Table 1.

fil-m	Filled pause with nasal sounds: The speaker hesitates and makes a nasal sound like *'mm'*.
fil-e	Filled pause with vocal sounds: The speaker hesitates and makes a a vocal sound like *'eeh'* or *'ouh'*.
spk-f	Speaker generated noise, breathing: The speaker is breathing loudly or making a fricative like sound.
spk-p	Speaker generated noise, sharp sound: The speaker makes a sound that is highly non-stationary like lip smack, coughing, clearing the throat, laughter or plosive like.
Sta	Stationary noise with approximately constant amplitude, like car noise or fan noise.
Int	Intermittent noise: Short non-stationary noise, like closing of a door, a phone ringing or rattling with paper.
<rep>	Repetition and correction. Example (fabricated): The patient has <has> bechterews <no> rheumatoid arthritis.
word word	Unpronounced words, including word fragments. Marked if it occur at the beginning or the end of the word.
**	Unintelligible stretches of speech.
word- -word	Signal truncation at the beginning or the end of the word. Caused by speaker or by technical disruption (cutting).

Table 1. Special symbols used in the proposed the Brage standard.

MOBELspon has a limited budget. Therefore we made some restrictions on the pronunciation dictionary. We only use phonetical transcription of the first occurrence of each new word heard on the recordings, thereby potentially missing multiple pronunciations. Medical terminology typically has long technical terms, often with Latin origin. The speakers use both a "Norwegian" accents and the original pronunciation, thus the pronunciations differ a lot. Examples: the word /*bechterews*/ is pronounced as /*b e k t @ r e v s*/, /*b e C t @ r e v s*/ and /*b e C t e r e v s*/, but only some of the alternative pronunciations were added to the dictionary, but no systematic attempt was made to do this thorough.

Furthermore, we only use the standard bokmål form of ordinary Norwegian words, which can be very different from the dialect words used by the speakers. A reason not to include dialects in the pronunciation dictionary is the

costly development of such a dictionary. It may be more practicable to use data driven learning methods with posterior knowledge of the complete edited patient record, to adaptively cope with the dialect of the speaker and thereby hopefully increasing the recognition accuracy for later use.

2.3 The speech corpus

MOBELspon contains a total of 150 minutes of spontaneous speech and 47.5 minutes of read speech, recorded from 5 female and 5 male speakers. All participants were medical students in their forth or fifth year of education, they were between 22 and 26 years old and had a wide diversity in the dialect background. See Table 3 for more details.

The vocabulary size is approximately 2600 words. The number of spoken words is approximately 16300 for spontaneous speech and 7700 for read speech, giving speech-rates of 92 and 145 spoken words per minute respectively. The spontaneous speech corpus has a much lower speech-rate, mainly because the speakers had to think of what they were going to say while dictating, thereby making long pauses.

The annotated speech corpus contains a high degree of para-linguistic sounds. A total of 7.52 % of the read speech and 11.6 % of the spontaneous speech corpus contain the non-words as explained in Table 1. The frequency of speaker generated noise types are given in Table 2. We can see that the breathing sound /*spk-f*/ is about equally prominent on both speaking styles. The main reason for this is the head mounted microphone used during recording of MOBELspon, because it captures loud breathing sounds from the speakers. As expected the spontaneous speech has more of the other types of speaker generated noise.

Noise type	Spont. speech	Read speech
spk-f	4.38	4.50
spk-p	1.88	0.93
fil-e	3.64	0.41
fil-m	0.54	0.04

Table 2. Occurrence in percent of the different speaker generated noise types in MOBELspon.

By listening we can hear that some of the words in the spontaneous speech often have a pronunciation of words that is worse than the read speech. This come in addition to the dialect diversity only heard in the spontaneous speech corpus.

3. RECOGNITION SYSTEM

A set of recognition experiments were done using the HTK program package [9]. The feature vectors consisted of 13 mel-frequency cepstral coefficients (including 0th cepstral coefficient) and their first and second derivatives. These were computed from a 25 ms Hamming window and updated every 10 ms. The acoustic modeling was done by a set of 5517 tied-state triphone models identical to those described in [5]. Each model has a three-state left-to-right structure with 16 Gaussian mixtures components in each state. A three state HMM is also used to model silence, but in addition we utilized a single state HMM to model short pauses between words.

Speaker adaptation was done for each speaker, by using supervised Maximum Likelihood Linear Regression (MLLR) [10] with a binary regression class tree with 128 leaf nodes. The number of regression classes used is dependent on the amount of adaptation data.

We did not have any acoustic models for the four different types of speaker generated noise sounds for Norwegian, so we made new ones by taking the models of those phonemes that had the most likely acoustic similarities as a starting point. Specifically we used the phonemes /*f*/, /*p*/, /2:/ and /*m*/ to model the speaker generated noise sounds /*spk-f*/, /*spk-p*/, /*fil-e*/ and /*fil-m*/, respectively. Ongoing research has improved these models by training them with both supervised and unsupervised methods. These results will be presented in a future paper.

For these initial experiments we used a simple word loop language model that consisted of all the word occurring in MOBELspon. Many pronunciation variations have been observed in the database due to the different dialect background of the speakers and the existence of less common medical terms.

4. PRELIMINARY RESULTS

This section presents results from all speakers. We divided our data into a test set and an adaptation set. The adaptation set was about 7 minutes for each speaker, where 1/3 was read speech and 2/3 spontaneous speech. The rest of the available data were used for testing.

Table 3 shows the recognition results for the different combinations of speakers, type of speech and acoustic models. Note that adaptation is based on a combination of read and spontaneous speech.

As reported in [5], the speaker independent acoustic models used in the recognition system were based on south-east dialect. Only M2 and F2 come from this area. They had also the highest recognition accuracy for read speech without adaptation. All the other speakers had a

dialect background that is different from normalized Norwegian. Thus, their speech was accented, and it was not surprising that they performed poorly with speaker-independent models. However, after speaker-adaptation of the read speech they got a higher relative improvement of the recognition.

Speaker	Dialect	RS	RSA	SS	SSA
M1	Lillehammer	31.28	47.33	34.76	45.78
M2	Oslo	54.42	63.47	14.19	38.11
M3	Haugesund	22.82	48.55	0.20	38.18
M4	Værdal	48.83	53.91	27.67	36.83
M5	Trondheim	42.69	50.72	20.93	34.45
F1	Sandnes	2.01	35.18	2.69	31.18
F2	Oslo	52.94	59.19	34.70	57.87
F3	Bodø	30.65	58.87	14.81	38.86
F4	Nordheim-sund	40.27	68.49	19.09	42.53
F5	Bergen	35.48	58.42	5.92	38.43
Total		36.14	54.41	17.50	40.22

Table 3. Word accuracy [%] for the different speakers and speaking styles. RS = read speech, RSA = read speech adapted, SS = spontaneous speech and SSA = spontaneous speech adapted. M = male and F = female speaker.

For spontaneous speech without adaptation the results were poor. This is due to the fact that both the dictation style and the dialect of most of the speakers are mismatched with respect to both the acoustic models and some of the words in the pronunciations dictionary. MLLR adaptation gave a consistent improvement for all speakers.

5. CONCLUSIONS

MOBELspon is a newly developed research database of spontaneous speech for Norwegian. An annotation standard to cope with spontaneous speech features is tested on this database. Preliminary recognition results show for example only slightly above 40 % word accuracy for adapted spontaneous speech using a word loop language model. This result is fare below acceptable performance for real systems.

We expect that we will get a substantial improvement by using a (bigram) statistical language model. However, presently we do not have access to such a language model for this application. Further, techniques to increase recognition accuracy, as more advanced models to handle the spontaneous speech features and data driven methods to capture dialect words, would hopefully increase the performance.

ACKNOWLEDGMENT

We would like to thank the phonetician Prof. Arne Kjell Foldvik for his work on transcribing the MOBELspon database, and PhD. Ingunn Amdal and PhD. Knut Kvale for proposing the new annotation standard.

REFERENCES

[1] G. Rigoll, “An overview on european projects related to spontaneous speech recognition,” ISCA and IEEE Workshop of Spontaneous Speech Processing and Recognition, Tokyo Japan, april 2003.

[2] S. Furui, “Recent advances in spontaneous speech recognition and understanding,” ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo Japan, Apr. 2003.

[3] BRAGE:, ”Brukergrensesnitt med naturlig tale” Available: http://www.tele.ntnu.no/projects/brage/

[4] MOBEL:, ”Mobil elektronisk pasientjournal” Available: http://mobel.digimed.no/

[5] T. Holter, E. Harborg, M. H. Johnsen, and T. Svendsen, “ASR-based subtitling of live TV-programs for the hearing impaired,” in Proc. Int. Conf. on Spoken Language Processing (ICSLP), vol. 1, (Beijin, China), Oct. 2000.

[6] H. Lærum, G. Ellingsen, and A. Faxvaag, “Doctors’ use of electronic medical records systems in hospitals: cross sectional survey,” British Medical Journal 2001;323;1344-1348, vol. 323, pp. 1344—1348, 2001.

[7] The Norwegian SpeechDat(II) database. Avalible: http://www.telenor.no/fou/prosjekter/taletek/speechdat

[8] SAMPA Phonetic alphabet for Norwegian, Available: www.phon.ucl.ac.uk/home/sampa/norweg.htm

[9] S. Young, et al, The HTK Book version 3.2. Cambridge University, Dec. 2002.

[10] Legetter C. J. Woodland P. C., “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.” Computer Speech and Language, vol. 9. no. 2. 1995, pp. 171-185.