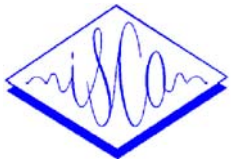

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ON-LINE CAPTIONING OF TV-PROGRAMS FOR THE HEARING IMPAIRED

Erik Harborg[1], Trym Holter[1], Magne Hallstein Johnsen[2], Torbjørn Svendsen[2]
[1] SINTEF Telecom and Informatics
[2] Norwegian University of Science and Technology
N-7465 Trondheim, Norway
Erik.Harborg@informatics.sintef.no
http://www.informatics.sintef.no/ and http://www.tele.ntnu.no/

ABSTRACT

A system for on-line generation of closed captions for broadcast of live TV-programs is described. During broadcast, a commentator formulates a possibly condensed, but semantically correct version of the original speech. These compressed phrases are recognized by a continuous speech recognizer, and the resulting captions are directly fed into the teletext system.

This application will provide the hearing impaired with an option to read captions for live broadcast programs, i.e., when off-line captioning is not feasible.

The main advantage in using a speech recognizer rather than a stenography-based system (e.g., Velotype) is the relaxed requirements for operator training. Also, the amount of text generated by a system based on stenography tends to be large, thus making it harder to read.

Keywords: Continuous speech recognition, large vocabulary, speaker adaptation, on-line captioning, aids for the hearing impaired.

1. INTRODUCTION

In this paper we describe our current efforts in developing a system for generation of closed captions for broadcast of live TV-programs, based on automatic speech recognition (ASR).

This application is aimed to be a new service from the Norwegian Broadcasting Corporation (NRK), offering the hearing impaired a possibility to enjoy live broadcast TV-programs. Today, this service is not offered by NRK.

In some countries, systems based on stenography (like e.g., Velotype) have been employed for this task. These systems enable the skilled operator to write text at a normal speaking rate. However, NRK has chosen not to implement such a system mainly due to the fact that the operators need a long training time to obtain the required speed. Also, the complexity of the stenography equipment generally limits the operator to produce verbatim transcriptions. Thus, the amount of generated text tends to be large, making it difficult for the users to read the captions before new ones show up on the screen.

In a system based on ASR, our preliminary experience shows that the commentator to a larger degree can concentrate on extracting the verbal program content. However, this is still not a trivial task, as the operator should simultaneously speak the sentences as well as being attentive to what is said next in the program. This is not for everyone to handle, but some people deal with this situation with very little training.

Our system is presently under development, and we will in this paper provide a functional description of it. We will discuss the particular problems encountered during the development of the system, how they have been solved, and also present some preliminary results on the performance of the system.

2. SYSTEM OVERVIEW

Figure 1 shows how the system will be operating. A commentator located in a separate studio is watching the currently transmitted TV-program on a screen. Equipped with headphones and a microphone, the commentator relays the verbal information of the program, either verbatim or in condensed form. The comments are automatically recognized by a continuous speech recognizer and fed into the existing teletext system. The users select the optional captioning through the teletext system.

The ASR-based captioning system contains two main modes, i.e., the user enrolment mode and the commentator mode. The main component in the user enrolment mode is the speaker adaptation procedure. To provide the necessary data, new operators are asked to record a set of utterances. Previously registered users may log directly into the commentator mode, which also contains an option to modify the vocabulary.

At present the language model and vocabulary are adapted to programs within the category news & politics and include about 12K words. This requires that the vocabulary can be easily changed on a daily basis, depending on what is presently in the news. The operator, who is not required to have any particular knowledge in speech technology, should perform this task. Thus, the

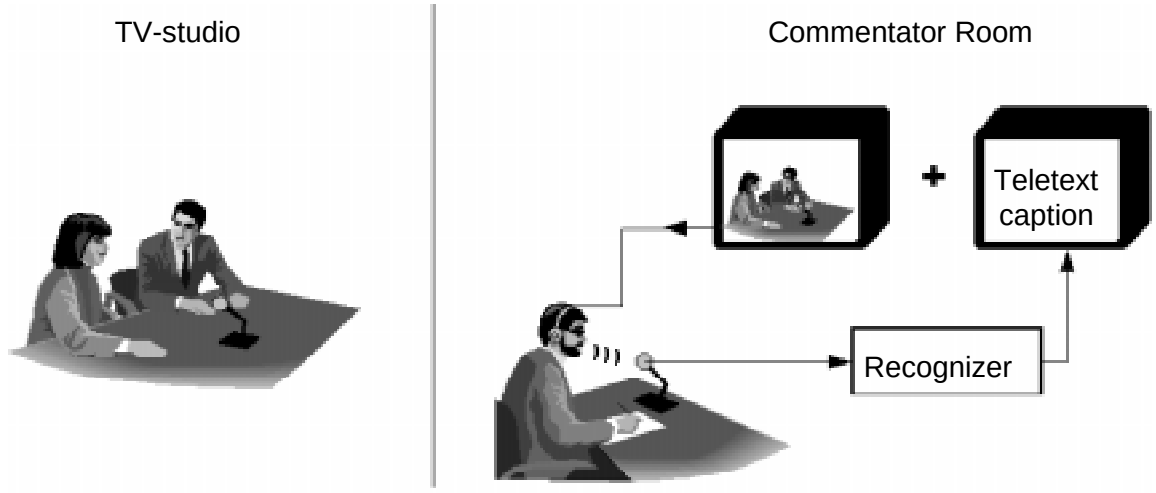

Figure 1. Operation of the on-line captioning system.

system must provide the user with transcriptions of the words that are added to the vocabulary.

In this first version of the system we have not put any efforts into making a dialect independent system. Therefore, the commentators are asked to speak in a normalized, Norwegian language.

If required, the broadcasting company may delay the broadcast of the TV-program with some seconds, in order to bring the program content and captions into synchronism.

At present, the system is running on a UNIX-based workstation. However, the final application will most probably be ported to a PC-based system running Linux or Windows NT.

3. DETAILED SYSTEM DESCRIPTION

Hidden Markov models with continuous mixture Gaussian densities are employed for acoustic modelling. We have used the Hidden Markov Model Toolkit (HTK) [1] during the development of the system. Additional parts have been programmed in the C/C++ programming languages.

3.1 Databases for text and speech

The recorded speech databases contain a total of about 17 hours of speech, uttered by 26 different speakers. This corresponds to about 12300 phrases containing about 120K words. 20 speakers are used for speaker independent training, while the remaining 6 are used for speaker adaptation and testing. The databases contain two different speech modes; read text and spontaneous speech. The spontaneous speech has been recorded in a setting that resembles a realistic situation for the particular application: The speaker was presented for a TV-program (without captions), and was asked to provide comments suitable for captioning for that particular program.

The basic training set covers almost 8 hours of speech (approximately 3 hours spontaneous and 5 hours read) for the 20 speakers, for a total of approximately 5200 phrases.

The testset constitutes a total of about 3.5 hours of spontaneous speech for the 6 test speakers, for a total of 2700 phrases. Up to 800 additional phrases (about an hour of both spontaneous and read text) are available for adaptation for each speaker in the testset.

For language modelling (see Section 3.6), we have access to a database which contains approximately 1 million words, collected from captioned news programs. This text has been edited in a semi-automatic manner in order to remove spelling errors, dialect words, abbreviations, special characters, etc.

3.2 Acoustic preprocessing

The basic feature vector consists of 13 mel-frequency cepstral coefficients (MFCCs, 0th order coefficient included), extended by 1st and 2nd order derivatives, i.e., a total of 39 coefficients. These are computed from a 25 ms Hamming window and updated every 10 ms.

3.3 Acoustic modelling

The acoustic modelling procedure is similar to the tutorial example of the HTK Book [1]. For context independent modelling, we have utilised the 46 symbols in the Norwegian SAMPA phonetic alphabet [2] plus additional schwa, silence (all units modelled by three state left-to-right HMMs), and tee (modelled by a single-state HMM) symbols. Rather than flat start initialisation, initial models have been created from the Norwegian EUROM.0 [3] and EUROM.1 databases [4].

From the context independent models, the training proceeded by building word-internal context dependent models for all triphones occurring in the training set. Phonetic decision tree clustering was then employed in order to reduce the total number of distinct states from 25695 to 1153, thus improving the generalisation abili-

ties. Currently we model the observation density in each state by a 12-component Gaussian mixture pdf with diagonal covariance matrices.

3.4 Static and dynamic dictionaries

The complete dictionary consists of two parts. The main partition will cover many of the common words used in the Norwegian language. However, it would not be possible to include initially all specialized words, proper nouns etc., which would be required for the application in a fixed dictionary. Therefore, we will add a dynamic, user-editable dictionary, which can be changed on a daily basis. The words "president" and "Christmas" are typical examples from the static and dynamic vocabulary, respectively, as the latter is a news topic for only a short time every year.

By spelling as well as pronouncing a new word, a phonetic transcription of the word is automatically proposed, which may then be accepted or edited by the operator. The proposed transcription is generated in a Viterbi decoding procedure, i.e., the best sequence of phones for that particular word is determined. When all editions are performed, the language model must be automatically updated. This is briefly discussed in Section 3.6.

3.5 Speaker adaptation

Adaptation for each commentator is facilitated using Maximum Likelihood Linear Regression (MLLR) [5]. A maximum of 800 phrases have been used for adaptation. During adaptation, we intended to let all triphones sharing the same centre phone constitute one regression class. However, merging was needed in order to avoid classes with very few training tokens. The number of regression classes was thus reduced from 49 to 44.

3.6 Language modelling

For these initial experiments, we have designed various bigram language models. These have been compared to simple unigram and 0-gram language models. At this stage we have not exploited the text database, but extracted the N-gram statistics from the annotated speech databases within a closed vocabulary framework.

In addition to a standard, backed-off word-based bigram, we have developed a smoothed bigram model based on word-classes [6]. The word-classes were obtained through a text corpus analysis performed using a grammatical tagger [7]. After the most infrequent classes had been merged with similar classes, approximately 130 different grammatical classes were left. As we have access to a very limited text corpus, class-based bigrams may improve the generalisation abilities compared to a standard bigram. In addition to this, a bigram based on grammatical classes may offer advantages in the language model updating needed in accordance with changes in the dynamic vocabulary. In this case, the operator can be asked to classify the new word into one of a small number of grammatical classes. However, as will be seen in Section 4, the current performance of the class-based bigram is not satisfactory.

3.7 Graphical user interface

The graphical user interface is presently based on the X window system. The interface contains the following features:

Registration of new users: By supplying a userID, a directory for that user is generated, where all the speaker specific files will be stored (adapted models, user specific pronunciations, speech files, etc.).

Speaker adaptation: The operator is prompted to read a number of phrases, with the option to accept or reject the latest recording. After all sentences have been recorded, model adaptation is automatically performed, and the user is allowed to enter the "commentator"-mode.

Editing the dynamic vocabulary: The operator is prompted to both enter the spelling and pronunce the new word. A phonetic transcription is proposed for the user, who in turn accepts or modifies it. The complete dynamic vocabulary can be presented in a scrollable window. Entries can be deleted or modified. When finished, the language model must be automatically updated.

Present status of the system: A window shows the present status of the system (i.e., "Recording", "Model Adaptation", etc.).

Commentator mode: The recognized text is presented in a separate window. The captions will also be written to an output port of the computer for use by a separate teletext system.

4. PRELIMINARY RESULTS

In this section, we report results from experiments performed with the 6 different speakers in the test set. The number of test phrases per speaker ranges from 274 to 401 in the experiments below.

In this early stage, focus has been attached to design of sub-word acoustic models. In Table 1, a series of experiments performed with speaker-independent as well as speaker-adapted models are reported. As would be expected, the MLLR adaptation consistently improves the recognition rates for all speakers. We note that the largest improvements are found for speakers 3 and 4, who are the only speakers in the test set who are not from southeastern Norway. Their dialect background is very different from normalized Norwegian. Thus, their normalized speech is accented, and it is not surprising that they perform poorly with speaker-independent models. However, after speaker-adaptation they achieve recognition rates comparable to the average of the whole group.

A number of different language models have been tested, as reported in Table 1. As would be expected,

Table 1. Recognition rates for the six test speakers.

Adapted	Grammar	%Correct words, speaker no.						
		1	2	3	4	5	6	Avg.
No	Bigram	79	79	70	73	84	84	79
Yes	Bigram	85	88	82	90	85	90	87
Yes	Word-class bigram	64	71	61	69	66	68	67
Yes	Unigram	62	69	59	66	64	65	64
Yes	Free grammar	51	56	47	53	51	53	52

the backed-off word bigram outperforms the others. The class-based bigram performs significantly worse than the word bigram, and only slightly better than the unigram. This indicates that the class-bigram itself is too smooth, thus failing to capture information relevant to the recognition task.

5. FUTURE WORK

We are currently performing new recordings to extend the training database. In particular, the number of speakers needs to be increased. However, we feel confident that in this next stage, we need to focus mainly on language modelling. We will take the text database into account and possibly increase its size. Still, the text resources are very limitied, and we will for this reason keep on investigating class-based bigrams, possibly based on data-driven clustering. The heuristics for language model updating in accordance with the dynamic vocabulary will also be of uttermost importance for the final result. This heuristics will of course be highly dependent on which language model we choose to implement in the final stage.

In parallel with this activity we seek to improve the response time of our current decoder in order to meet the demands for a small delay for the captioning of live broadcast TV-programs. We also believe that the pronunciation modelling should be improved, possibly by adding alternate baseforms for words that exhibit large inter-speaker variation.

6. CONCLUSIONS

We have described our current efforts in developing a system for generation of closed captions for live broadcast TV-programs, based on Norwegian ASR. A functional description of the system has been given, and we have discussed some problems encountered during the development. Some preliminary results have also been reported, indicating that MLLR adaptation can compensate for the relatively large dialectal diversity in Norway, given that the operators are instructed to talk in a normalized manner. The results do also indicate that statistical language modelling is the major challenge in the next stage of our development project.

ACKNOWLEDGEMENTS

This work has been funded by the Norwegian Broadcasting Corporation (NRK), and inititated by Rolleiv Solholm, Managing Editor, NRK Teletext.

Also, NRK has performed the recordings of the speech data and the collection of the text database.

The tagger used to generate the word classes, has been developed by the University in Oslo and Lingsoft Inc.

REFERENCES

[1] Young S.J. et al.: *The HTK Book, version 2.1*, Cambridge University, March 1997.

[2] Kvale K., Foldvik A.K.: "Manual Segmentation and Labelling of Continuous Speech," in *Proc. ESCA Workshop on "Phonetics and Phonology of Speaking Styles: Reduction and Elaboration in Speech Communication,"* pp. 37.1-37.5, Barcelona, 1991.

[3] Grice M., Barry W.J., Fourcon A.: "Specification of EUROM0 assessment," Appendix B: Part 2 in *Support available from SAM-project for other ESPRIT speech and language work*, SAM-document G001/b/2, 1989.

[4] Sherwood T., Fuller H.: "Guide to EUROM.1 Speech Database", SAM-NPL-102, April 1992.

[5] 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.

[6] Witschel P., Hoege H.: "Experiments in Adaptation of Language Models for Commercial Applications," in *Proc. Eurospeech'97*, Rhodes, Greece, pp. 1967-1970.

[7] Hagen K., Johannessen J.B., Nøklestad A.: "A Constraint-Based tagger for Norwegian," in *Proc. XVI Scandinavian Conference of Linguistics*, 1999.