

# ASR-BASED SUBTITLING OF LIVE TV-PROGRAMS FOR THE HEARING IMPAIRED

*Trym Holter<sup>1</sup>, Erik Harborg<sup>1</sup>, Magne Hallstein Johnsen<sup>2</sup>, Torbjørn Svendsen<sup>2</sup>*

<sup>1</sup>SINTEF Telecom and Informatics, N-7465 Trondheim, Norway

<sup>2</sup>Norwegian University of Science and Technology (NTNU), N-7491 Trondheim, Norway  
Trym.Holter@informatics.sintef.no Erik.Harborg@informatics.sintef.no

## ABSTRACT

A system for on-line generation of closed captions (subtitles) 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 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 commentator training. Also, the amount of text generated by a system based on stenography tends to be large, thus making it harder to read.

## 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) [1] [2]. Here, we extend the work described in [3].

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. Nevertheless, a fair share of the people testing the

concept dealt satisfactorily with this situation after a very short training period.

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 the latest 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 recognised by a continuous speech recogniser and fed into the existing teletext system. The users select the optional captioning through the teletext system.

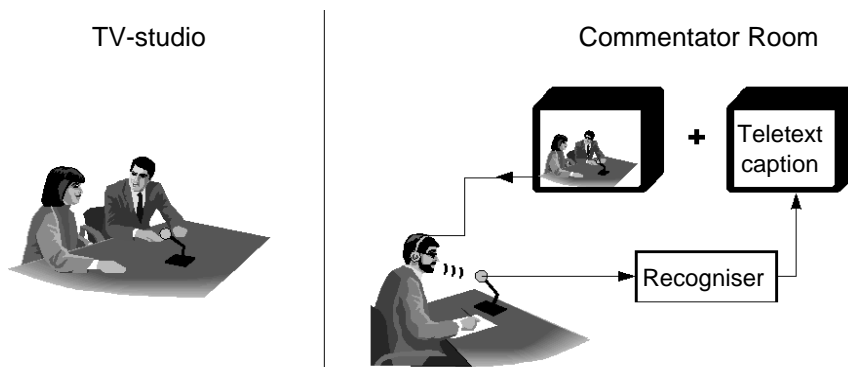
The ASR-based captioning system contains two main modes, i.e., the user enrolment mode and the commentator mode. The main component in the user enrolment mode is the speaker adaptation procedure. To provide the necessary data, new operators are asked to read a set of utterances. Previously registered 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 and politics, and include about 15K 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 system must provide the user with transcriptions of the new words that are added to the vocabulary.

At present we have not put any efforts into making a dialect independent system. Therefore, the commentators are asked to speak in a normalised Norwegian language.

## 3. DETAILED SYSTEM DESCRIPTION

Hidden Markov models with continuous Gaussian mixture densities are employed for acoustic modelling. We have used the Hidden Markov Model Toolkit (HTK) [4] during the development of the system.



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

### 3.1. Databases for Text and Speech

In our previous experiments [3], we had available a recorded speech database containing a total of about 17 hours of speech, uttered by 26 different speakers. We have now completed the recordings of the speech database with 30 additional speakers.

The complete speech database now contains about 19300 phrases with almost 190K words. This adds up to a total of about 28 hours of speech. 50 speakers are used for speaker independent training, while the remaining 6 are used for speaker adaptation and testing. The database contains 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 19 hours of speech (approximately 7 hours spontaneous and 12 hours read) for the 50 speakers, for a total of approximately 12200 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 used a database which contains approximately 1 million words, collected from captioned news programs.

### 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 [4]. For context independent modelling, we have utilised the 46 symbols in the Norwegian SAMPA phonetic alphabet [5] 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 [6] and EUROM.1 databases [7].

From the context independent models, the training proceeded by building word-internal context dependent models for all tri-phones occurring in the training set. Phonetic decision tree clustering was then employed in order to reduce the total number of distinct states from 35481 to 2000, thus improving the generalisation abilities. 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 covers 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 in order to include the "hot" words in the present news picture.

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 completed, the language model must be automatically updated.

### 3.5. Speaker Adaptation

Adaptation for each commentator is facilitated using Maximum Likelihood Linear Regression (MLLR) [8] and maximum a-posteriori (MAP) [9] techniques. MLLR makes use of a

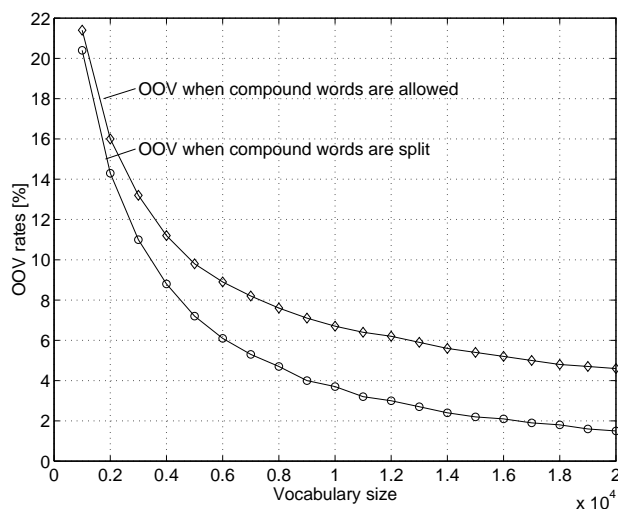
regression class tree to group the Gaussian mixture components, so that the set of transformations can be chosen according to the amount of available adaptation data. When large amounts of data are available, the MLLR-adapted HMMs can be further refined through MAP adaptation, as this scheme is specifically defined at the component level. In these experiments, a maximum of 800 phrases have been used for adaptation for each speaker.

### 3.6. Language Modelling

In our previous experiments [3] we designed various N-gram language models based on the annotated speech databases within a closed vocabulary framework. In this paper, we address the more realistic situation with an open vocabulary. N-gram statistics and vocabulary are extracted from a separate 1 million words text database within the news domain. For the present experiments, we have designed a backed-off, word-based bigram language model.

One major challenge we are facing is the large proportion of compound words that is found in Norwegian, as well as other languages including German, Dutch, and Swedish [10]. In particular, compound nouns formed by concatenation of single nominals occur frequently. As a consequence, for a given vocabulary size, the out-of-vocabulary (OOV) rate will usually be high for these languages, compared to e.g., English. To handle this problem, we have split the compound words in the text database as well as the annotated text associated with the speech test database. In Figure 2, the OOV rates are plotted versus vocabulary size before and after the splitting. For each vocabulary size  $L$ , the  $L$  most frequent words in the text database constitute the vocabulary. The OOV rate is estimated from the annotated test speech database.

We have not yet decided a strategy for merging words output by the speech recognizer in order to restore the compound words. However, for our task this is not believed to be of crucial importance, as the splitting of compound words will in most cases not change the meaning of the sentence.



**Figure 2:** OOV rates vs. vocabulary size before and after compound word splitting.

For these experiments we have designed a bigram language model for a 15K vocabulary extracted from the split version of the text database. The corresponding OOV rate is 2.2%. The bigram is mainly word-based, except for the proper nouns. These have been split into 6 classes; person names, geographical names, other names (typically companies and organisations), and the genitival versions of those three types. Within each class a uniform probability distribution is assumed.

## 4. EXPERIMENTS AND 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. As described in Section 3.6 we are using an open, 15K words vocabulary, decomposition of compound words, and a backed-off word bigram. The OOV rate for these experiments is 2.2%. In contrast, our previously reported results [3] was obtained using a closed vocabulary of 12K words, and a bigram extracted from the 120K words annotated speech database, which implies that the test data was applied in the language model design. However, the test set is exactly the same. It should be noted that this is not read text, but speech of a somewhat spontaneous nature.

In Table 1 and 2, a series of experiments performed with speaker-independent as well as speaker-adapted models are reported. Speakers 1, 2, and 3 are males, while speakers 4, 5, and 6 are females. As would be expected, the MLLR adaptation consistently improves the recognition rates for all speakers. Also, additional MAP estimation helps on performance. We note that the largest improvements due to adaptation 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 normalised Norwegian. Thus, their normalised 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. Also, it should be noted that these two speakers perform worse on the extended training set without speaker adaptation. This can be explained by increased mismatch between the basic acoustic models and the speech from these accented speakers. All the new 30 speakers in the extended training set are from southeastern Norway. However, speaker no. 4 turns out to be the overall best after speaker adaptation, probably because of her clear, distinct voice, and fluent speech with very little hesitation.

It is observed from Table 1 and 2 that the effect of just increasing the basic training set from 8 hours of speech (“old” model) to about 19 hours of speech (“new” model) is rather limited. For the non-adapted models this increases the average word recognition rate with about 3%. We feel that there is a potential for further improvement by better utilisation of the new training data. This will be explored in the future.

For the best speaker, almost 83% of the words are correctly recognised. For this application, this is a more relevant metric than the average result, both because we expect the performance to improve with more training of the operators, and because the operators who achieve best results will be preferred. As observed from Table 2, for this speaker about 32% of all sentences are correctly recognised. In order to reveal the usefulness of the

| Adapted  | Model | %Correct words, speaker no. |      |      |      |      |      |      |
|----------|-------|-----------------------------|------|------|------|------|------|------|
|          |       | 1                           | 2    | 3    | 4    | 5    | 6    | Avg. |
| No       | Old   | 61.3                        | 63.7 | 54.4 | 55.6 | 69.7 | 66.2 | 61.8 |
| No       | New   | 68.1                        | 68.5 | 54.1 | 53.0 | 72.5 | 71.2 | 64.6 |
| MLLR     | New   | 74.9                        | 79.1 | 70.6 | 81.1 | 75.8 | 77.7 | 76.5 |
| MLLR+MAP | New   | 75.6                        | 80.0 | 72.0 | 82.7 | 76.5 | 78.3 | 77.5 |
| MLLR+MAP | Old   | 74.2                        | 78.1 | 71.1 | 80.6 | 74.8 | 76.6 | 75.9 |

**Table 1:** Word recognition rates for the six test speakers.

| Adapted  | Model | %Correct sentences, speaker no. |      |      |      |      |      |      |
|----------|-------|---------------------------------|------|------|------|------|------|------|
|          |       | 1                               | 2    | 3    | 4    | 5    | 6    | Avg. |
| No       | Old   | 10.7                            | 8.9  | 6.9  | 6.5  | 14.2 | 15.9 | 10.5 |
| No       | New   | 14.5                            | 10.0 | 7.3  | 5.2  | 16.2 | 20.7 | 12.3 |
| MLLR     | New   | 20.2                            | 22.5 | 17.2 | 28.7 | 18.4 | 28.4 | 22.6 |
| MLLR+MAP | New   | 22.0                            | 23.9 | 18.3 | 32.3 | 19.1 | 31.1 | 24.5 |
| MLLR+MAP | Old   | 18.2                            | 21.7 | 18.6 | 29.7 | 19.5 | 25.8 | 22.3 |

**Table 2:** Sentence recognition rates for the six test speakers.

sentences containing errors, we have performed a subjective evaluation. It turns out that 32% of all the sentences are understandable, even though they contain one or more errors, 26% of the sentences are partially understandable, i.e. they contain some information, and the remaining 10% contain no information at all.

## 5. FUTURE WORK

In this next stage, we will focus mainly on language modelling. Due to the very limited amount of task specific text data, we want to investigate an approach based on interpolation with a language model trained with text collected from a different, but related domain. 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 more 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 results have also been reported, indicating that MLLR and MAP adaptation can compensate for the relatively large dialectal diversity

in Norway, given that the operators are instructed to talk in a normalised 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). Also, NRK has performed the recordings of the speech data and the collection of the text database. Telenor R&D supplied the text-to-phoneme generator used during pronunciation lexicon design.

## REFERENCES

- [1] Furui S., Zhang Z.-P., Ohtsuki K.: "On-line incremental speaker adaptation for broadcast news transcriptions," proc. IEEE ASRU Workshop, Keystone, USA, December 1999.
- [2] Imai T. et al. "Progressive 2-pass decoder for real-time broadcast news captioning," proc. ICASSP-2000, Turkey, June 2000.
- [3] Harborg E., Holter T., Johnsen M.H., Svendsen T.: "On-line Captioning of TV-Programs for the Hearing Impaired," proc. EUROSPEECH-99, Budapest, Hungary, September 1999.
- [4] Young S.J. et al.: *The HTK Book, version 2.2*, Cambridge University, January 1999.
- [5] 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.
- [6] 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.
- [7] Sherwood T., Fuller H.: "Guide to EUROM.1 Speech Database," SAM-NPL-102, April 1992.
- [8] 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.
- [9] Lee C.-H., Lin C.-H., Juang B.-H.: "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, 1991, pp. 806-814.
- [10] Carter D., Kaja J., Neumeyer L., Rayner M., Weng F., Wiren M.: "Handling Compound Nouns in a Swedish Speech-Understanding System," in *Proc. ICSLP '96, Philadelphia, USA*, pp. 26-29.