LIVE TV SUBTITLING
Fast 2-pass LVCSR System for Online Subtitling

Aleš Pražák, Luděk Müller
SpeechTech s.r.o., Morseova 5, 301 00 Plzeň, Czech Republic

J. V. Psutka, J. Psutka
Department of Cybernetics, University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic

Keywords: ASR, LVCSR, HMM, real-time, class-based language model, live TV, online subtitling.

Abstract: The paper describes a fast 2-pass large vocabulary continuous speech recognition (LVCSR) system for automatic online subtitling of live TV programs. The proposed system implementation can be used for direct recognition of TV program audio channel or recognition of a shadow speaker who re-speaks the original audio channel. The first part of this paper focuses on preparation of an adaptive language model for TV programs, where person names are specific for each subtitling session and have to be added to the recognition vocabulary. The second part outlines the recognition system conception for automatic online subtitling with vocabulary up to 150 000 words in real-time. The recognition system is based on Hidden Markov Models, lexical trees and bigram and quadgram language models in the first and second pass, respectively. Finally, experimental results from our project with the Czech Television are reported and discussed.

1 INTRODUCTION

There is a lot of hearing impaired people who have only limited access to the information contained in audio channel in the multimedia content. In the television - the most widespread mass media - there is an effort to access its multimedia content to these people with alternative textual information - the subtitles. Many public service televisions such as Czech Television have a duty by law to subtitle certain portion of their broadcasting. The subtitles should be added to all TV programs, even to the live TV programs with minimum delay. To meet this requirement the automatic speech recognition (ASR) technology is being introduced for live subtitling in some television companies, for example BBC (Evans, 2003).

Since the automatic speech recognition technology is not error-free and the recognition results are very dependent on the acoustic speech signal quality the trend is to use so-called shadow speaker who respeaks the original speech by his own words. The recognition system can be learned to professional speaker acoustic characteristics, manner of speech and even the vocabulary he or she uses. The recognition results are then more accurate so the subtitles can be well intelligible. In addition, the shadow speaker can simplify the original speech to meet the demands of the hearing impaired people for simple, easily readable subtitles. However, the training of the shadow speaker is a very time and money consuming process.

Some TV programs have clear acoustic speech signal and quite limited vocabulary so direct recognition of the TV program audio channel can be carried out with reasonable recognition accuracy. This approach can save expenses on shadow speakers and can be fully automated. So far we have prepared a system for automatic online subtitling of the live transmissions of the Czech Parliament meetings without use of a shadow speaker.

2 ADAPTIVE LANGUAGE MODEL

By law, the shorthand records of all Czech Parliament meetings are available for public use on the Internet. These shorthand records are amended to avoid slips of the tongue and to meet grammatical rules; however there is a huge amount of text from different elec-

Pražák A., Müller L., V. Psutka J. and Psutka J. (2007).
LIVE TV SUBTITLING - Fast 2-pass LVCSR System for Online Subtitling.
In *Proceedings of the Second International Conference on Signal Processing and Multimedia Applications*, pages 139-142
DOI: 10.5220/0002140301390142

toral periods to create a high quality language model. Unfortunately, this training text contains many non-standard (NS) words: abbreviations, acronyms, numbers written as figures, dates etc. Therefore the conversion of NS words to their standard forms (digit sequences, abbreviations and acronyms to the full word forms) is essential. This is called text normalization.

2.1 Text Normalization

The text normalization for languages with a low degree of inflection such as English is easier because the most conversions are unambiguous. However, in highly inflectional languages, such as Czech or other Slavic languages, one NS word can be converted to several standard forms, each of which has the same meaning but represents different morphological categories (gender, case and number). The morphological meaning of each NS word is given also by its context in a whole sentence. The method proposed in (J. Kanis, 2005) solves the task of finding the right standard form for a given NS word. This method is based on a tagger performing context-dependent morphological disambiguation of each word in a given sentence.

The text normalization system has two modules. The first module ensures NS word detection and classification and the second module the conversion itself. The detection and classification is based on regular expressions. We distinguish 17 different types of NS words, for example cardinal and ordinal number, date, abbreviation, currency, percentage etc. After the NS word is detected the second module converts it to the standard form. The conversion is algorithmic and uses the method proposed in (J. Kanis, 2005). Firstly, the NS word is converted to the basic form and then a tagger is used to find the morphological information which determines the right standard form.

The automatic NS word conversion accuracy is about 90 %. To improve this result we performed manual correction as postprocessing of the automatic conversion. The implementation of automatic text normalization with the manual correction accelerates process of the text normalization more then ten times in comparison with full manual text normalization.

2.2 Language Model Classes

Now a standard n-gram language model can be trained on the normalized text of the parliament meetings from three electoral periods. The language model contains the names of representatives (deputies and government members) from these three electoral periods. However, this language model does not allow subtitling of parliament meetings from different (including future) electoral periods, due to missing names of representatives elected in those periods. In addition, each name has been seen in different contexts in the training text, so the n-gram probability mass is split among many names of representatives. Even though the parliament meeting speeches contain only 1.5 % of representative names, there are some TV programs such as sport transmissions where the portion of names in the commentary exceeds 15 % and the described problem becomes crucial. Because the names of representatives are known before the first parliament meeting, the language model can be adapted before the actual subtitling. We created a language model in which each class incorporates one concrete Czech grammatical case of all representative names. The language model was trained in two steps.

Firstly, the names of the representatives from the electoral periods corresponding to the training text were automatically inflected to all grammatical cases. This inflection was based on specific rules for different cases and different inflectional patterns. Each name in the training text was then replaced by a tag representing the name's grammatical case. Unfortunately, different grammatical cases can have the same morphological form (especially female names), so manual classification of ambiguous morphological forms was still necessary.

In the second step, taking into account these tags instead of the individual names, a class-based n-gram language models were trained. Five classes have been created and filled by inflected names of representatives from demanded electoral period. The four classes contain multi-words `"first_name+surname"` and words `"surname"` in different grammatical cases. The class probabilities were split between these words in proportion to their frequency in the training text. The last fifth class contains reverse multi-words `"surname+first_name"` in the first grammatical case. Other forms of names did not occur in the training text.

Using an adaptive language model with classes for names, our online subtitling system can be used for any electoral period of parliament meetings. This approach can be effectively used also for other named entities, for example company or state names.

3 LVCSR SYSTEM

The fast 2-pass large vocabulary continuous speech recognition system developed at the Department of Cybernetics, University of West Bohemia is the main module of the whole subtitling system.

3.1 Acoustic Processing

The analogue input speech signal is digitized at 44.1 kHz sampling rate and 16-bit resolution format. The aim of the front-end processor is to convert continuous acoustic signal into a sequence of feature vectors. We performed experiments with MFCC and PLP parameterizations, see (J. Psutka, 2001) for methodology. The best results were achieved using 27 filters and 12 PLP cepstral coefficients with both delta and delta-delta sub-features. Feature vectors are computed at the rate of 100 frames per second.

Each individual basic speech unit is represented by a three-state HMM with a continuous output probability density function assigned to each state. In this task, we use only 8 mixtures of multivariate Gaussians for each state. The choice of an appropriate basic speech unit with respect to the recognition network structure and its decoding is discussed later.

3.2 Recognition Network

Our LVCSR system uses a lexical tree (phonetic prefix tree) structure for representation of acoustic baseforms of all words of the system vocabulary. In a lexical tree, the same initial portions of word phonetic transcriptions are shared. This can dramatically reduce the search space for a large vocabulary, especially for inflectional languages, such as Czech, with many words of the same word stem. The automatic phonetic transcription (with pronunciation exceptions defined separately) is applied to all words of the system vocabulary and resulted word baseforms for all pronunciation variants are added to the lexical tree.

To better model the pronunciation of words we used triphones (context dependent phonemes) as the basic speech units. By using a triphone lexical tree structure, the in-word triphone context can be easily implemented in the lexical tree. However, the full triphone cross-word context leads to fan-out implementation by generation of all cross-word context triphones for all tree leaves. This results in enormous memory requirements and vast computational demands. To respect the requirement of the real-time operation we have proposed an approximation of the triphone cross-word context.

One of the possible approaches is to use monophones (context independent phonemes) instead of triphones on the word boundaries. However, this brings the necessity to train two different types of acoustic model units and also mutually normalize the monophone and triphone likelihoods. To cope with this problem, we use only triphone state likelihoods and merge the triphone states corresponding to the same monophone within a given phone context. As the system vocabulary is limited, not all right and left cross-word contexts have to be modeled. This approach results in so-called biphones that represent merged triphone states with only one given context - right in the root and left in the leaves. The biphone likelihood is computed as the mean of the likelihoods of merged triphone states. The proposed biphone cross-word context represents a better approximation than a simple replacement of triphones by monophones on the word boundaries. In addition, this approach increases neither the recognition network complexity nor the decoding time, but only the duration of offline recognition network creation.

3.3 Recognition Network Decoder

Since bigram language model is implemented in the first pass, a lexical tree copy for each predecessor word is required. The lexical tree decoder uses a time-synchronous Viterbi search with token passing and effective beam pruning techniques applied to re-entrant copies of a lexical tree. The beam pruning is used inside and also at the level of the lexical tree copies, but a sudden increase of hypothesis log-likelihoods occurs due to application of language model probabilities at time of word to word (lexical tree to lexical tree) transitions. Fortunately, early application of the knowledge of language model can be carried out by factorizing language model probabilities along the lexical tree. In the lexical tree, more words share the same initial part of their phonetic transcriptions and thus only the maximum of their language model probabilities is implemented towards the root of the lexical tree during the factorization. In addition, commonly used linear transformation of language model log-likelihoods is carried out for optimal weighting of language and acoustic models.

To deal with requirement of real-time operation, an effective method for managing lexical tree copies is implemented. The algorithm controls lexical tree to lexical tree transitions and lexical tree copies creation/discarding. The number of lexical tree copies decoded in real-time is limited, so the control algorithm keeps only the most perspective hypotheses and avoids their undesirable alternations, which protects the decoding process from time consuming creation of lexical tree copies. The algorithm also manages and records tokens passed among lexical tree copies in order to identify the best path at the end of the decoding. In addition, for word graph generation not only the best, but several (n-best) word to word transitions are stored. HTK Standard Lattice Format (S. Young, 1999) is used to store the word graph.

Figure 1: Online subtitling of TV transmission of the Czech Parliament meeting.

In the second pass of the recognition system, the word graph is rescored with class-based 4-gram language model trained as described in section 2. To allow progressive subtitle displaying, the word graph creation and its rescoring can be performed even several times per second. The whole LVCSR system can effectively use multi-core computer systems, so the proposed fast 2-pass LVCSR system implementation handles tasks up to 150 000 words in real-time with the delay about one second.

4 EXPERIMENTS

The acoustic model for subtitling of the parliament meetings was trained on 40 hours of parliament speech records with manual transcription. We used 42 Czech phonemes. As the number of Czech triphones is too large, phonetic decision trees were used to tie their states. Now, subtitling of the Czech Parliament meetings works with 3 729 different HMM states of a speaker and gender independent acoustic model.

Three language models were trained on about 18M tokens of normalized Czech Parliament meeting transcriptions (Chamber of Deputies only) for the experiment. The first one is the baseline bigram back-off language model (BLM) with Good-Turing discounting trained directly from the training text, i.e. without name classes. The second one is an adapted class-based 2-gram language model (ALM2) for the first pass and the last one is an adapted class-based 4-gram language model (ALM4) for word graph rescoring in the second pass. The last two language models were trained from training text incorporating tags representing the name classes. The vocabulary size is almost 113 000 words. The SRI Language Modeling Toolkit (Stolcke, 2002) was used for training.

Table 1: Experimental results with baseline and adapted language models.

Language model	Test perplexity	Recognition correctness	Recognition accuracy
BLM	305	86.13 %	83.67 %
ALM2	292	87.02 %	84.42 %
ALM4	208	87.55 %	85.06 %

Five parliament speech records from different electoral period than the training text, half an hour each, were chosen for the testing. The OOV word rate is 1.51 % for BLM, 1.35 % for ALM2 and ALM4. It is important to notice that about 50 % of the OOV words are slips of the tongue. The perplexity and recognition results are reported in Table 1.

5 CONCLUSION

We designed our fast 2-pass LVCSR system implementation that is suitable for automatic online subtitling using original TV program audio channel or using a shadow speaker. Many word errors are caused only by missing prepositions and wrong endings of flexible words, so the subjective readability of automatically generated subtitles is very high.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education of the Czech Republic under project MŠMT 2C06020.

REFERENCES

Evans, M. J. (2003). Speech recognition in assisted and live subtitling for television. *BBC R&D White Paper*, 065.

J. Kanis, J. Zelinka, L. M. (2005). Automatic numbers normalization in inflectional languages. In *SPECOM 2005, 10th International Conference SPEECH and COMPUTER*.

J. Psutka, L. Müller, J. V. P. (2001). Comparison of mfcc and plp parameterization in the speaker independent continuous speech recognition task. In *EUROSPEECH 2001, 7th European Conference on Speech Communication and Technology*.

S. Young, e. a. (1999). *The HTK Book*. Entropic Inc.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. In *ICSLP 2002, 7th International Conference on Spoken Language Processing*.