Benefit of a Class-based Language Model for Real-time Closed-captioning of TV Ice-hockey Commentaries

Jan Hoidekr, J.V. Psutka, Aleš Pražák, Josef Psutka

Department of Cybernetics, University of West Bohemia in Pilsen
Pilsen, Czech Republic
{hoidekr, psutka_j, aprazak, psutka}@kky.zcu.cz

Abstract

This article describes the real-time speech recognition system for closed-captioning of TV ice-hockey commentaries. Automatic transcription of TV commentary accompanying an ice-hockey match is usually a hard task due to the spontaneous speech of a commentator put often into a very loud background noise created by the public, music, siren, drums, whistle, etc. Data for building this system was collected from 41 matches that were played during World Championships in years 2000, 2001, and 2002 and were transmitted by the Czech TV channels. The real-time closed-captioning system is based on the class-based language model designed after careful analysis of training data and OOV words in new (till now unseen) commentaries with the goal to decrease an OOV (Out-Of-Vocabulary) rate and increase recognition accuracy.

1. Introduction

The recognition of spontaneous speech like the ice-hockey commentary is a very hard task. A commentator usually speaks over the noise in a sport arena. The noise is very different in each match, sport arena and changes also during running match. In spite of the fact that a vocabulary created from an ice-hockey commentary seems to be relatively well limited we show that the lexicon formed by collecting words of training matches does not cover commentaries of new matches sufficiently and a high OOV rate brings about many recognition problems. First of all we tried to recognize ice-hockey commentaries using a "classic" approach, i. e. by collecting training data and building acoustic and language models using this data. Using speech data of real commentaries for acoustic modeling provided a good basis for automatic recognition of disturbed speech signal in hockey commentaries. The problem concerning the high OOV rate was solved after careful analysis of OOV words in new (till now unseen) matches. We found out that majority of them were the proper names of players, names related to nationalities of players/teams, names of towns etc., which were relatively specific for a commentary of a given match. Consider that more than 10% of words in the recognition lexicon are names of players appearing in many grammatical cases. It is probably caused by a high degree of inflection and derivation of the Czech language as a representative of the large family of Slavic languages. We tried to alleviate a high OOV rate by a suitable knowledge-based selection of the recognition lexicon. Our approach was based on the fact that the list of players is known before the match starts, so we can add their names to the recognition lexicon. We developed mechanism how to automatically derive all applicable grammatical forms of players' names indicated on line-ups and also some other categories mentioned above. The set of all possible forms of proper names and names of states and nationalities was used as a basis for building class-based language models. This approach decreased the OOV rate from 7.7% to 4.5% and increased the overall recognition accuracy more than by 9% assuring the accuracy of recognized names above 90%.

2. Baseline system

To build a baseline large vocabulary continuous speech recognition system for automatic closed-captioning of ice-hockey commentaries we collected and annotated training data from 41 matches that were played during World Championships and Winter Olympic Games in years 2000, 2001, and 2002 and were transmitted by the Czech TV channels. These matches were annotated with respect to specific acoustic events (background noise, music etc.) using special annotation software Transcriber 1.4.1. (Psutka et al., 2003). Annotated data was used to build both acoustic and language models, which were integrated into our real-time recognition system ERIS. For tests of developed system we randomly selected three commentaries of matches played during the World Championship in 2005.

2.1. Speech and language processing

Recognition experiments were performed with the recognition engine, which was built at the Department of Cybernetics, University of West Bohemia in co-operation with spin-off firm SpeechTech. The recognition engine is based on a statistical approach. It incorporates the front-end, acoustic model, language model and real-time decoder (ERIS). The basic speech unit of our system used for acoustic modeling is a triphone. Each individual triphone is represented by a three-state HMM with a continuous output probability density function assigned to each state. At present we use 8 mixtures of multivariate Gaussians for each state. As the number of Czech triphones is too large, phonetic decision trees were used to tie states of Czech triphones. Now our system works with 7k7 different states.

The digitization of an analogue signal was provided at 44.1 kHz sample rate and 16-bit resolution format. The aim of the front-end processor is to convert continuous speech into a sequence of feature vectors. Several tests were performed in order to determine the best parameterization of the acoustic data. We experimented with MFCC and PLP parameterizations. The best results were achieved using 27 filters and 16 PLP cepstral coefficients with both delta and delta-delta sub-features. Therefore one feature vector con-

tains 48 coefficients. Feature vectors were computed at a rate of 100 frames per second.
There were also tested many noise reduction techniques to reduce often very intense background noise, which came from audience and from the game itself. As the best noise-reduction technique in our case seems to be J-RASTA, for details see (Psutka et al., 2001).
The language models were created from 342,157 running words (tokens) obtained from manual transcripts of all training commentaries. The final size of the recognition lexicon was 22,392 words. The first experiments were performed using the bigram back-off language model with Good-Turing discounting. We used SRILM toolkit (Stoclke, 2002) to compute language model statistics.

2.2. Real-time decoder

The LVCSR system with a real-time decoder uses the lexical tree (phonetic prefix tree) structure for representation of acoustic baseforms of all words in the system vocabulary. In a lexical tree the same initial portions of word phonetic transcriptions are shared. This can dramatically reduce 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 is applied to all words in the system vocabulary and resulted word baseforms for all pronunciation variants are added to the lexical tree. The phonetic transcription of foreign words and phonetic exceptions were defined separately.
In addition, the voice assimilation phenomenon (Pražák et al., 2003) encountered in the Slavic languages was also taken into account. Due to the cross-word voice assimilation one or more last phonemes of a word can be influenced by one or more initial phonemes of the successor word. Acoustic baseforms of all words were adapted to both the voiced and the unvoiced right cross-word contexts. These new adapted baseforms were added to the lexical tree.
Since the bigram language model considerably reducing task perplexity is used, a lexical tree copy for each predecessor word is required. The trigram language model which requires more lexical tree copies (for each two predecessor words) does not lead to improvements adequate to its more complex implementation.
The lexical tree decoder uses a time-synchronous Viterbi search with token passing and effective beam pruning techniques applied to re-entrant copies of the lexical tree. To deal with the requirement of the real-time operation, an effective method for managing lexical tree copies is implemented. The algorithm also manages and records tokens (comprising time indexes and scores) passed among lexical tree copies in order to identify the best path at the end of the decoding.

2.3. Application framework

The implementation of speech signal acquisition and subtitles displaying is platform dependent. Our implementation on Microsoft Windows system is based on Microsoft DirectX technology (Linetsky, 2001).
The application framework consists of two DirectShow filters one-way connected to each other. Thus there are two detached DirectShow filters for audio and video stream. This solution is independent of the source media and can be used for digital records via cascade of DirectShow filters as well as for TV signal acquired by Windows Driver Model (WDM) capturing filters. The speech signal is passed directly to the LVCSR system engine running at the separate thread. The recognized word sequence is forwarded via system pipes to the subtitle displayer implemented as a video DirectShow filter. Recognized subtitles are incorporated into the source video stream by our video DirectShow filter.

2.4. Lexical statistics

The Czech language belongs to Slavic languages which demonstrate a high degree of inflection. It means that there are more word forms for one basic word. It causes a large amount of word items in a recognition lexicon and brings about problems with collection of enough text data to train corresponding language models. The ice-hockey commentary is spontaneous speech with many broken words, fillers and incomplete sentences. In Figure 1 you can see how the lexicon of the training matches covers commentaries of three whole test matches (the curve of no. 1). We successively compared the cumulative lexicons (lexicons arisen collecting words of the first n training matches) with a sequence of words present in particular test matches. It is evident from Figure 1 that the OOV rate decreases up to 7.7% which is still relatively very high level if we consider that this result was obtained using lexicon built from commentaries of more than 40 matches, which contained 342,157 running words.

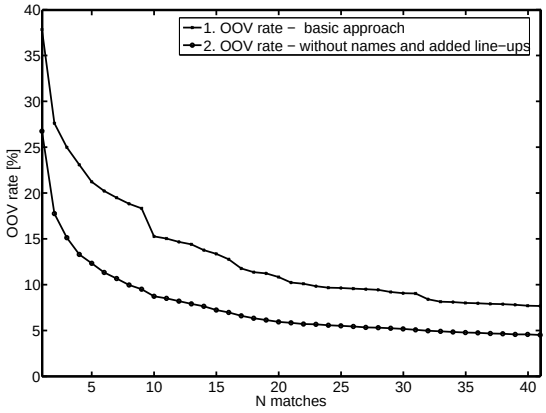

Figure 1: OOV rates with growing lexicon

2.5. Recognition results of baseline system

For tests of developed system we selected commentaries of matches played during the World Championship in 2005. There were randomly chosen nine 5-minute segments from 3 matches (one 5-minute segment within a hockey match period) containing in total 3,729 running words. To compare our recognition system based on the real-time decoder ERIS we performed parallel tests also on AT&T decoder (Mohri et al., 2000), which works more than ten times slower than real-time. Table 1 shows results of recognition

experiments for both decoders given in standard measures of Correctness (Corr [%]) and Accuracy (Acc [%]). For each 5-minute test segment the Table 1 also contains the OOV rate (O_r [%]) and perplexity (PPL). It is evident that the AT&T decoder gave slightly better results in comparison with the real-time decoder ERIS. This can be explained by the necessity of an intensive pruning of search space owing to real-time function of ERIS.

dec.	**AT&T**		**ERIS**			
	Corr	Acc	Corr	Acc	O_r	PPL
T1_1	62.67	55.11	58.89	52.22	10.9	554
T1_2	81.43	72.86	78.10	70.48	5.0	561
T1_3	68.98	57.49	64.97	58.82	10.7	663
T2_1	68.88	57.44	62.47	54.00	7.8	650
T2_2	77.37	60.89	71.79	54.47	6.1	728
T2_3	74.21	63.32	67.62	57.59	5.4	594
T3_1	66.05	54.40	58.49	49.90	4.8	496
T3_2	67.61	54.08	66.48	55.21	7.3	608
T3_3	65.19	56.54	58.15	51.31	8.5	428
Avg	**69.23**	**58.27**	**64.71**	**55.73**	**7.4**	**587**

Table 1: Recognition results – baseline system

3. Adaptive language models

Analyzing recognition errors of the first set of experiments we found out that many errors were caused by a high level of OOV rate. A large portion of OOV words usually consisted of new names of players, who were active in the given match. Because the list of players is usually known before a match starts we could try to predict changes and modify both the recognition lexicon and language model. The idea was to mark manually all names of players in the lexicon and replace this group of words by the hypothetic word $< NAME >$. Let us note that we had nearly 17% of names in our original lexicon which consisted of 18,611 different words. Then in accordance with an actual hockey match we added to the lexicon all new names using known line-ups of both teams.

3.1. Automatic generation of inflected word forms

As it was said above the Czech language has a high degree of inflection, so the one word can have many grammatical forms. Because the names in line-ups are in the basic form we have to create for them all possible grammatical forms. It means to generate for every name up to 20 new forms in Czech language. We developed the technique for automatic derivation of all name forms in the given list of players. Of course, a very important step of this technique is phonetic transcription of each word item. Because the players came from the whole world the phonetic transcription of their names was solved by a rule-based approach in combination with a very large pronunciation lexicon of foreign names. Let us note that the inflection of the foreign names is derived in Czech from the basic phonetic form. As this technique is not absolutely perfect a manual revision and correction is needed but this process is very fast and effortless.

3.2. Word-class of player names

Our first attempt to adapt the recognition lexicon was based on using only one word class for all names introduced on line-ups of the given match. We created this class with uniform probability distribution for all individual names. The class of names contained about 800 words in average for each new match, so the recognition lexicons for individual test matches were different. This class of names was added to the recognition lexicon and incorporated also to the language model through the general word $< NAME >$. After adding the class of names to the vocabulary we got smaller lexicon (in comparison with the lexicon containing all names and collected from commentaries of all training matches) but this lexicon covered test commentaries much better. So, by addition of list of names (in all grammatical forms) to the recognition lexicon, the OOV rate decreased from 7.7% to 4.5% (see the curve of no.2 in Figure 1). Let us mention also the perplexity of an adapted language model which is similar as this in the baseline system (see Table 1). It could indicate not so good probability estimation of names inside a class.

dec	**AT&T**		**ERIS**			
	Corr	Acc	Corr	Acc	O_r	PPL
T1_1	67.11	66.22	62.22	58.22	3.8	573
T1_2	84.05	76.90	77.86	70.95	1.7	565
T1_3	70.32	63.90	66.58	63.10	5.1	729
T2_1	68.42	57.67	66.13	60.18	5.9	649
T2_2	76.26	62.85	72.07	60.61	5.0	575
T2_3	73.93	65.33	69.34	61.60	4.3	550
T3_1	69.33	60.74	60.74	54.19	4.5	483
T3_2	70.70	58.87	71.83	63.10	3.9	677
T3_3	67.81	60.16	60.76	56.34	4.4	481
Avg	**71.74**	**63.07**	**67.02**	**60.61**	**4.3**	**587**

Table 2: Recognition results – word-class of names

3.3. Multiple word-classes of names according to grammatical cases

Establishing word-class of players' names brought significant improvement of the recognition accuracy. However after analysis of recognition results we found out that many errors still remained on the level of names of players. The names were often recognized in a bad grammatical form mainly with incorrect endings. This was probably caused by the fact that a name has usually very similar pronunciation (the same stem) in all grammatical cases. We decided to split the word-class of names into 10 new classes according to their grammatical forms. The first four classes were prepared for basic substantive in different cases (the Czech language knows 7 cases), four classes were assigned to possessive names, one class to first names of players and one to special forms of names. As special items of each class we created compound words that were put together from the first name and surname of individual players. We found out that this approach sometimes improves recognition of name-surname combinations and avoids the bad insertion of the first name in front of a given surname. We

also grouped names in different grammatical cases especially when they had identical written forms and pronunciations. Recognition results provided by this solution are shown in Table 3. We can observe significant improvement of the system functionality. It is interesting that the recognition accuracy obtained only for pronounced names exceeded 90%.

dec	AT&T		ERIS			
	Corr	Acc	Corr	Acc	O_r	PPL
T1_1	70.67	67.56	66.67	62.89	3.9	417
T1_2	87.14	80.24	81.43	75.71	1.7	399
T1_3	75.13	68.45	67.65	64.44	5.2	526
T2_1	69.34	59.27	66.13	60.87	6.5	534
T2_2	77.93	65.36	75.98	64.80	4.5	447
T2_3	75.36	65.62	69.91	63.04	4.6	455
T3_1	70.55	62.37	62.99	57.26	4.8	387
T3_2	73.80	61.69	74.93	66.48	4.0	457
T3_3	69.82	62.37	63.78	59.96	4.6	389
Avg	**74.12**	**65.78**	**69.48**	**63.66**	**4.4**	**446**

Table 3: Recognition results – multiple word-classes

3.4. Word classes of nationalities and states

Another category of interesting words is the group of names of states and nationalities whose players and referees appeared in matches. Because the Czech TV transmitted more than one half of matches with the Czech team we were afraid that language model statistics could be slightly deviated mainly when the commentary should follow the match of two foreign teams (not the Czech team). We approached this problem in the same way as in the case of personal names. We automatically derived a set of words in all grammatical cases for each name of state and nationality. During creation of individual classes we also distinguished nationalities of rivals in the given match and players of others nationalities mentioned in the commentary. We could observe an interesting phenomenon in commentaries of matches of the Czech team. In these commentaries there was lower relative frequency of words like "Czech Republic, Czech, ... " in comparison with names of states not playing this match. The commentator usually uses words "our team, we, ..." and the commentary and corresponding lexicon were for these matches slightly different. So we decided to divide matches into two groups, with and without the Czech team. These modifications brought a little but very important improvements of recognition results. We substantially reduced the recognition errors of words which brought the meaning of nationalities and states. These words are very important for readability of subtitles. The results of this improvement are given in Table 4.

dec	AT&T		ERIS			
	Corr	Acc	Corr	Acc	O_r	PPL
T1_1	71.56	68.44	66.67	62.89	3.7	406
T1_2	86.67	79.29	81.62	76.19	1.7	409
T1_3	75.67	68.98	67.65	64.71	5.2	485
T2_1	69.57	59.04	66.36	61.33	6.3	568
T2_2	78.49	65.08	76.82	66.48	4.2	497
T2_3	75.36	65.90	69.91	63.04	4.3	460
T3_1	70.76	62.37	63.80	58.28	4.8	393
T3_2	74.08	62.98	74.93	66.20	4.0	506
T3_3	70.02	61.97	63.98	60.56	4.4	406
Avg	**74.39**	**65.92**	**69.75**	**64.15**	**4.3**	**458**

Table 4: Recognition results – nationalities and states

4. Conclusions

We presented the real-time speech recognition system for automatic closed-captioning of TV ice-hockey commentaries. The performance of this system was based on adaptive class-based language models. We showed that a "blind" collection of training data with the goal to use it for building the recognition lexicon and language models isn't useful due to accumulation of large number of personal names in the lexicon. The addition of new names from actual line-ups to the lexicon together with introduction of class-based language model and division of classes according to grammatical cases increased significantly the recognition accuracy and rapidly improved readability of generated subtitles (mainly owing to the high accuracy of recognized names (above 90%)). Although the overall recognition accuracy isn't still sufficiently high we found that the final closed-captioning system is acceptable for people who have hearing defects.

5. Acknowledgment

Support for this work was provided by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1ET101470416.

6. References

M. Linetsky, 2001. *Programming Microsoft DirectShow*.

M. Mohri, M. Riley, and F. Perceira, 2000. *Weighted Finite-State Transducers in Speech Recognition.* Intern. Workshop on Automatic Speech Recognition.

A. Pražák, F. Jurčíček, L. Müller, and J. V. Psutka. 2003. Voice assimilation phenomenon and its implementation in lvcsr system with lexical tree and bigram language model. In *3rd International Conference on Signal, Speech and Image Processing*.

A. Pražák, L. Müller L., and J. V. Psutka. 2005. Lvcsr system for automatic online subtitling. In *10th Intl. Conference SPEECH and COMPUTER*.

J. Psutka, L. Müller, and Psutka J. V. 2001. Comparison of mfcc and plp parameterization in the speaker independent continuous speech recognition task. In *Proc. of Eurospeech 2001*.

J. Psutka, J.V. Psutka, P. Ircing, and J. Hoidekr. 2003. Recognition of spontaneously pronounced tv ice-hockey commentary. In *Proc. of the ISCA&IEEE Workshop on Spontaneous Speech Processing and Recognition*.

A. Stoclke. 2002. Srilm - an extensible language modeling toolkit. In *Proc. Intl. Conf. Spoken Language Processing.*