FOUR-PHASE RE-SPEAKER TRAINING SYSTEM

Aleš Pražák, Zdeněk Loose, Josef Psutka, Vlasta Radová
Department of Cybernetics, University of West Bohemia, Plzeň, Czech Republic

Luděk Müller
SpeechTech s.r.o., Plzeň, Czech Republic

Keywords: Speech recognition, LVCSR, Online captioning, Re-speaker training, Application.

Abstract: Since the re-speaker approach to the automatic captioning of TV broadcastings using large vocabulary continuous speech recognition (LVCSR) is on the increase, there is also a growing demand for training systems that would allow new speakers to learn the procedure. This paper describes a specially designed re-speaker training system that provides gradual four-phase tutoring process with quantitative indicators of a trainee progress to enable faster (and thus cheaper) training of the re-speakers. The performance evaluation of three re-speakers who were trained on the proposed system is also reported.

1 INTRODUCTION

With the rise of computer technology and the progress in the probabilistic approach to the large vocabulary continuous speech recognition (LVCSR), this technology suggests itself to be used for automatic or semi-automatic captioning of live TV broadcasting. There are in general two ways how to use the speech recognition for live TV captioning. The first one is a direct recognition of the audio stream of a TV program. This approach is usable only for very specific TV programs with defined acoustic characteristics, minimum non-speech sounds, limited domain of the discourse and a specific manner of speech. A typical task for this approach is fully automatic captioning of parliament meetings or broadcast news. The parliament meeting procedures as well as broadcast news design in almost all countries ensure stable acoustic environment and cultivated speech of only one person at a time, so the recognition accuracy can be high enough for trouble-free reading and understanding of the captions (Neto et al. 2008). We operate such a system together with the Czech Television, the public service broadcaster in the Czech Republic, for almost four years (Pražák et al. 2007).

The second approach for semi-automatic captioning of live TV broadcasting uses so-called "re-speaker" (or shadow-speaker) technology. The re-speaker is a skilled and specifically trained speaker, who listens to and re-speaks the original dialogues of a TV program, alternatively using his/her own words. This approach is suitable for arbitrary TV programs, especially for programs with more speakers speaking simultaneously or with a noisy acoustic environment, such as TV debates or sport programs. It also simplifies the task of LVCSR significantly – the re-speaker works in a quiet environment, uses a well-defined acoustic channel and produces a refined speech. Moreover, the acoustic model in the LVCSR can be personalized specifically for the given speaker. Alternatively, the re-speaker is allowed to use his/her own words, so the final captions can be shorter (easily readable) and more comprehensible for hearing-impaired. This in fact represents the translation from one language to the same one.

Probably the first broadcasting company that introduced LVCSR technology in the real caption generation process was BBC in 2003 (Evans 2003). Since then, similar systems have been developed and employed in production use in several countries all around the world (Boulianne et al. 2006), (Homma et al. 2008).

2 TRAINING SYSTEM

It is impractical to bring a re-speaker to the real

Pražák A., Loose Z., Psutka J., Radová V. and Müller L..
FOUR-PHASE RE-SPEAKER TRAINING SYSTEM.
DOI: 10.5220/0003604502170220
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2011), pages 217-220
ISBN: 978-989-8425-72-0

captioning system and let him/her do the real job of a skilled re-speaker throwing away the resulting captions for a few months. More effective is to develop a training system that shortens (and thus cheapens) the training process. We have developed a special training system for re-speakers that provides gradual training process under surveillance of a skilled supervisor and with quantitative indicators of their progress to enable easier and more objective decision about their suitability for re-speaking.

2.1 Overview

The proposed training system is a multi-user system that uses real-time LVCSR system to create captions upon the re-speaker's dictation and keyboard commands. We use in-house LVCSR system that is based on Hidden Markov Models (HMMs), lexical (phonetic prefix) trees and a trigram language model. The implementation is focused on low-latency real-time operation with very large vocabularies on multi-core systems. Due to high efficient decoder parallelization and a graphic processor unit (GPU) utilization, we can recognize more than 500 000 words in real-time on four-core notebooks. This is very important to allow intensive re-speaker training in all conditions with demanded recognition accuracy. The system supports also word graph generation for confidence measure computation.

After the system start, the re-speaker chooses his/her profile and automatically sets the microphone volume. This is accomplished in two steps - in the first step, the volume on silence is set (to filter the background noise), in the second step the optimal volume of speech is set. The soundtrack of a TV program is played during the whole process to simulate real training conditions that influence the re-speaker's utterance.

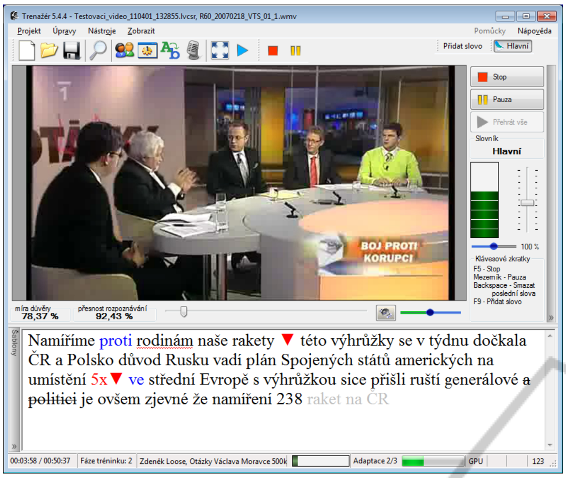
Now, the re-speaker chooses one of four training phases based on his/her training progress. The first training phase enables training of a re-speaker’s skill to listen and speak simultaneously. The second phase assists in optimizing the re-speaker’s utterance to the LVCSR system demands. The third training phase enables free re-speaking with a support of some keyboard commands and finally, the fourth phase simulates the real captioning system with all the features such as manual punctuation.

2.2 Training Phases

The first training phase is intended to train re-speaker’s skill to listen and speak simultaneously. The re-speaker opens prepared video file and practices speaking while playing any part of the video. The aim is not to re-speak word by word, but become accustomed to speaking meaningfully while listening to and perceiving the original soundtrack. This phase does not employ LVCSR system, but all utterances are recorded for later playback by the re-speaker. Each time the re-speaker starts playing video, new record is created and metadata, such as timestamp, position in video and repetition count of each segment, are logged. This allows the supervisor to trace re-speaker’s training process. Recorded utterances are played back simultaneously with the video as they were recorded, but the sound volume balance between recorded utterances and the original soundtrack can be set at will.

The second phase of the training system assists in optimizing the re-speaker’s utterance to the LVCSR system demands, so this phase integrates the LVCSR system and displays its output to the re-speaker. The main objective of the re-speaker is to re-speak utterances in the original soundtrack word by word so that the recognition accuracy is as high as possible. The re-speaker just mechanically re-speaks what he/she listens to, so he/she can focus on altering the utterance (mainly the pronunciation) and its influence on the recognition results. This implies that the video files for the second phase should contain neither overlapping speakers nor slips of the tongue nor out-of-vocabulary (OOV) words. Anyway, the speech rate of some speakers can be too high for some re-speakers at the beginning of the training. That is why there is an option to slow down the playback rate of the video file in real-time. The WSOLA algorithm ensures that the pitch remains the same even for high changes of the playback rate (Verhelst 2000).

We have implemented two quantitative indicators that indicate the training progress of the re-speaker. The first one is common recognition accuracy, but supplemented with highlighting of misrecognitions in the recognized text. A Levenshtein alignment is carried out just during recognition in real-time, so that the recognition substitutions (underlining), insertions (strikeout) and deletions (omission triangle) may be highlighted (see Figure 1). By definition, the video file transcription has to be available for the second phase.



míra důvěry 78,37 %

přesnost rozpoznávání 92,43 %

Namíříme proti rodinám naše rakety ▼ této výhrůžky se v týdnu dočkala ČR a Polsko důvod Rusku vadí plán Spojených států amerických na umístění 5x▼ ve střední Evropě s výhrůžkou sice přišli ruští generálové ~~a politici~~ je ovšem zjevné že namíření 238 raket na ČR

00:03:58 / 00:50:37 | Fáze tréninku: 2 | Zdeněk Loose, Otázky Václava Moravce 500k | Adaptace 2/3

Figure 1: Phase two of the training system.

The second indicator of the training system is a so-called "suitability measure" with the meaning of suitability of the re-speaker's utterance to the LVCSR acoustic model. The suitability measure is computed only from correctly recognized words and the language model is omitted, so it expresses the re-speaker's ability to comply with the LVCSR system almost independently from the particular recognized text. In the same way as in the case of misrecognitions, the words with low suitability measure are highlighted (different word color) in the recognized text (see Figure 1). During the playback of the recorded utterances, the words just being played are highlighted (different background color), so the relationship between played sound and misrecognitions and suitability measure of words can be tracked. This indicator should have increasing tendency during the whole training process of the re-speaker.

The next important task of the second phase of the training system is the acoustic data gathering and acoustic model adaptation to the speaker's voice characteristics. The acoustic model adaptation is applied in two stages. The first adaptation stage is applied iteratively just during the recognition in real-time, so the recognized text can be improved immediately. An unsupervised incremental fMLLR adaptation is carried out in the background, so the re-speaker's effort is not influenced (Pražák et al. 2009). In the second adaptation stage, all the data gathered during the whole re-speaker's training are used for MAP adaptation enhanced with the SAT approach. The acoustic model adapted in that manner is then used in the real captioning system.

The third phase of the training system is very close to the real captioning system, in order to make the re-speaker ready to the real captioning. The re-speaker re-speaks video files with his/her own words and learns to use some of the features that improve the final captions. The recognition accuracy indicator is not available, since the re-speaker‘s utterance transcriptions cannot be known in advance, whereas the suitability measure is still displayed to check the re-speaker's training progress. As the suitability measure should be computed only from the correctly recognized words, the confidence measure estimated from the LVCSR system result is used to guess the correctly recognized words instead of exact transcription.

It comes from the principle of the LVCSR system, that last few words of the recognized text change as new acoustic signal is received and the best hypothesis based on acoustic and language model is recomputed. Since only static closed captions are still required in the TV broadcasting, these last so-called "pending" words (four at maximum) are ignored during the caption generation, but they can be displayed to the re-speaker highlighted (different word color), so he/she knows, which words can be eventually corrected (see Figure 1). To be able to quickly correct any of the pending words (not only the last one), the best method is to erase them all and re-speak consequently. The best way is to use idle re-speaker‘s hands and keyboard commands, so the re-speaker presses optional key and re-speaks erased words fluently (Wald et al. 2007). This feature should not be ignored, because it can dramatically decrease the error rate of the final captions.

The next feature that allows generation of high-quality captions during the real captioning session is the possibility to dispatch the pending words to the captioner to be broadcasted immediately. This is very important when the re-speaker does not speak for a longer time (because of TV jingle or he/she is listening ahead) and the pending words remain unsent. It can significantly reduce the delay between the words uttered in the original soundtrack and corresponding words in the captions.

The fourth phase of the training system simulates the real captioning with all the features needed. In addition to the erasing or dispatching of the pending words a re-speaker should use his/her hands to make punctuation and for other special functions. The system for punctuation indication is closely connected with the LVCSR system. The key pressed during the inter-word pause is processed by the system that presents the punctuation symbol directly in its result. This approach allows the LVCSR system to benefit from extra information using language model considering punctuation symbols as words. Similar method is used for speaker change marking too. Other keys are reserved for special announcements and situations.

The re-speaker is trained on the real video files containing OOV words (mainly named entities) that should be added to the LVCSR system just during the captioning. This is a crucial procedure that should not take much time. We have implemented a method for word additions in real-time during the recognition of LVCSR system. The simplest way is to type a word (or multi-word) and confirm it. The system searches for the word in the special large lists of named entities and if succeeds, it adds the word with correct phonetic transcription(s) to the LVCSR system. In the case of unknown word, automatic phonetic transcription is proposed and optionally modified by the re-speaker. This approach minimizes the time for word addition in most cases.

3 EVALUATION

To assess the training possibilities of the described system, we have evaluated three re-speakers who got through all the training phases. The re-speakers produced the captions for the same real TV debate (61 minutes) and the final captions were evaluated. The recognition accuracy (including the pending words corrections) and some statistics collected during the captioning are presented in Table 1.

Table 1: Evaluation of re-speakers on the TV debate.

	NZ	**EK**	**PZ**
Training time	138 hours	98 hours	78 hours
Recognition accuracy	97,37 %	97,34 %	94,32 %
Suitability measure	85,66 %	84,41 %	82,38 %
Words	4294	2960	4217
Word additions	8	12	12
Word corrections	83	66	73
Word dispatches	60	136	100
Commas	319	170	287
Full stops	341	248	335
Question marks	49	48	57
New speakers	109	103	120

4 CONCLUSIONS

The proposed system facilitates the re-speaker training by decomposing the education process into four gradual phases, making the training easier and thus faster. The overall training time is highly individual, but according to our expertise, we expect the time of intensive training plan to be set from 2 to 3 months (100 training hours at minimum).

As the development of our training system continues, we want to allow synchronous training of caption correctors that are indispensable for error-free captioning of some critical TV broadcastings.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Education of the Czech Republic under the projects MŠMT 2C06020 and MŠMT LC536 and by the grant of The University of West Bohemia, project No. SGS-2010-054.

REFERENCES

Boulianne, G., Beaumont, J.-F., Boisvert, M., Brousseau, J., Cardinal, P., Chapdelaine, C., Comeau, M., Ouellet, P., Osterrath, F., 2006. In *International Conference on Spoken Language Processing*.

Evans, M. J., 2003. Speech Recognition in Assisted and Live Subtitling for Television. *WHP 065*. BBC R&D White Papers.

Homma, S., Kobayashi, A., Oku, T., Sato, S., Imai, T., Takagi, T., 2008. New Real-Time Closed-Captioning System for Japanese Broadcast News Programs. In *Computers Helping People with Special Needs*. Springer.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., Caseiro, D., 2008. Broadcast news subtitling system in Portuguese. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Pražák, A., Müller, L., Psutka, J. V., Psutka, J., 2007. LIVE TV SUBTITLING - Fast 2-pass LVCSR System for Online Subtitling. In *International Conference on Signal Processing and Multimedia Applications*.

Pražák, A., Zajíc, Z., Machlica, L., Psutka, J. V., 2009. Fast Speaker Adaptation in Automatic Online Subtitling. In *International Conference on Signal Processing and Multimedia Applications*.

Verhelst, W., 2000. Overlap-add methods for time-scaling of speech. In *Speech Communication*. Elsevier.

Wald, M. and Bell, J.-M. and Boulain, P. and Doody, K. and Gerrard, J., 2007. Correcting automatic speech recognition captioning errors in real time. In *International Journal of Speech Technology*.