AN AUTOMATIC CAPTIONING SYSTEM FOR TELEMEDICINE

Y. Zhao, X. Zhang, R-S. Hu, J. Xue, X. Li, L. Che, R. Hu, L. Schopp*
Department of Computer Science, University of Missouri, Columbia MO 65211 , USA
*Department of Health Science Management
zhaoy@missouri.edu, {xz934, rhe02, jxwr7, xlm7b, lc7y9, rhq2c}@mizzou.edu, schoppl@health.missouri.edu*

ABSTRACT

In this paper, we present a first exposition of an automatic closed captioning system designed to assist hearing impaired users in telemedicine. This system automatically separates telehealth conversation speech between a health care provider and a client into two streams and provides real-time captions of health care provider's speech to client. The captioning system is based on the state-of-the-art technology of large vocabulary conversational speech recognition, encompassing speech stream separation, acoustic modeling, language modeling, real-time decoding, confidence annotation, and human-computer interface, with innovations made in several components. The system currently handles a vocabulary size over 46 K. Real-time captioning performance at the average word accuracy of 77.95% is reported.

1. INTRODUCTION

Telemedicine or telehealth (videoconferencing for health care) has opened a world of specialty health services to persons who are otherwise unable to access appropriate care. Despite enormous public investment in telehealth systems, people who are deaf or hard of hearing encounter serious barriers to access current systems because of unsatisfactory audio quality, audio/video delay, and limited sign language and lip reading capabilities due to video distortion of motion scenes. Users with hearing loss in general rate captioned video materials valuable [1], and captioning is widely used in television broadcast programs [2]. If users with hearing loss are to benefit from telehealth, captioning system must be used. However, unlike captions used in television programs and educational tape materials that are generated by human expertise, it is unfeasible that a human expert be employed in each telehealth session to deliver closed captions, due to both factors of cost and availability. Therefore, developing a voice-driven captioning system for telehealth by utilizing the state-of-the-art automatic speech recognition technology is of great significance to the access of telehealth by the deaf and hard of hearing users.

Automatically transcribing broadcast news and lecture speech by using spoken language technology have been actively studied in recent years. DARPA Hub4's leading efforts since 1990's played a major role in launching large vocabulary continuous speech recognition (LVCSR) research in broadcast news in the United States [3]. Despite of wide variations in speaking styles, accents, and environmental conditions, LVCSR performance has been rapidly improved and achieved operational capability. Automatic broadcast news captioning was developed and put in use by NHK in Japan, where accuracy greater than 95% was achieved [4]. Recently, automatic transcription of lectures or presentations is drawing significant attentions in LVCSR [5]. Compared with broadcast news, lectures typically have a higher variation in speaking style, fluency, environmental conditions, and less constrained in syntax. Relatively little supervised training data also poses challenges in both acoustic and language modeling. Recent research in [5] showed an error rate of 32.4% on the Translation English Database (TED) task.

Automatic captioning for telemedicine bears similarities and differences from the above tasks. In the current system, focus is made on captioning health care provider's speech alone due to two reasons. First, health care providers regularly use telehealth systems and therefore it is potentially feasible to collect individual provider's speech to train accurate acoustic models. Second, the need for captioning is primarily on the clients' side since there is generally a much higher percentage of clients with hearing loss than providers. Since conversation in telehealth is carried out in alternating turns of provider and client, preprocessing is needed to extract provider's speech stream out of the conversation for subsequent recognition. Health care providers' speech are spontaneous, have varying degrees of filled pauses, repetitions, repairs, etc. Each session of telehealth conversation may focus on one medical specialty problem or cover multiple specialties. To facilitate conversation understanding, it is important that the captions are delivered in real-time and with low latency. On the other hand, captioning word rate needs to be appropriate (≤ 150 words/minute) to avoid excessive amounts of captions that frustrate slow readers whose primary language is American Sign Language. Furthermore, it is important that medical terms be captioned correctly, and the system should provide necessary functions beyond automatic speech recognition for detection and correction of mistakes.

The captioning system as described in this paper has fulfilled most of the above described functions, while a few interface functions are currently under development. The system was trained by speech data collected in telehealth and five medical doctors have served as health care providers. An average captioning word accuracy of 77.95% has been achieved, with confidence annotation accuracy of 84.74%. The system achieved real-time captioning on TigerEngine v1.1, a one-pass speech decoding engine developed for the captioning system in the authors' laboratory [6].

This paper is organized into five sections. In Section 2, the procedure of data collection for the captioning task is explained. In Section 3, several key components of the captioning system are described, including speech stream separation, acoustic modeling, language modeling, real-time decoding, confidence annotation, and user interface. Experimental results are provided in Section 4, and a conclusion is made in Section 5.

This work is supported in part by National Institutes of Health under the grant NIH 1 R01 DC04340-01 A2.

2. TELEMEDICINE DATA COLLECTION

The University of Missouri Telemedicine Network was used as the site for data collection. Seven health care providers participated in the data collection. Recordings were made in sessions, with one session for one client, and each session lasted for about 20~30 minutes. Conversation topics were primarily on neuropsychology, internal medicine, and dermatology. Health care providers' speech data were automatically extracted from conversation speech and organized into records that were separated by pauses (see Section 3.1). There were about 51 hours of data extracted from the recordings, with about 24 hours of data from the seven health care providers. The speech data of health care providers were transcribed by experienced personnel, with a total of 305818 words and 8.02% words being medical terms.

In choosing microphone for speech recording, a tradeoff between recoding quality and unobtrusiveness was made. High quality recording is desired for accurate captioning, but close-talking microphone with a signal-to-noise ratio (SNR) of about 40 dB is undesirable since having a microphone close to mouth would interfere with lip reading, which is often needed by people with hearing loss. A wireless microphone system was chosen, where a lapel microphone was pinned to the cloths, which was unobtrusive to lip reading by hearing impaired clients, non-tethering to provider, and yielded a SNR of 25 dB that was higher and more consistent than SNR of far field microphone used in telehealth. Sampling rate of 16 KHz was used in recording.

3. MAJOR SYSTEM COMPONENTS

The captioning system for telehealth with its major system components is shown in Fig. 1. Doctor and patient conversation speech is recorded and doctor's speech stream is extracted from the conversation. Automatic speech recognition system delivers recognized words to the user interface module, where words are automatically annotated by confidence values, word rates/minute is estimated based on recognized words and duration of underlying speech to provide feedback to doctor for speech rate adjustment, and a pen edit function is provided to doctor to make selective corrections on caption word blocks. Doctor controls the timing of sending corrected captions to patient.

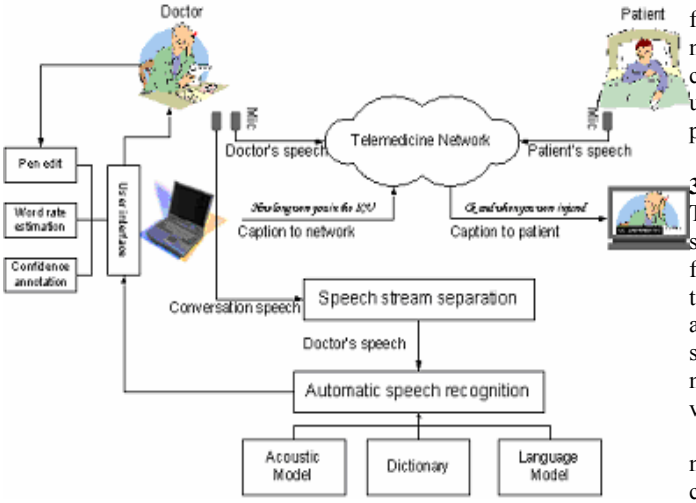

Fig. 1 Proposed automatic captioning system for telehealth.

3.1. Speech stream separation

Two methods were investigated for separating speech streams of health care provider and client. One is based on statistical speaker identification with online adaptation of speaker models [7], and the other is based on echo-cancellation capability of teleconferencing, where the former uses single channel recording and the latter uses dual channels. The echo-cancellation based method is adopted in the currently described system due to its higher reliability, even though average error rates of the two methods are comparable. The dual channel recording works in the following way. On doctor's site, wireless microphone acquires one channel input of speech conversation z, where provider's speech is directly recorded and client's speech come from a loud speaker mounted on the wall in provider's room. The second channel input is from teleconferencing audio output of patient's speech y. Sliding window of 0.1 sec. is applied to both recording channels in synchrony. For a window W(t), energy levels in the two channels, Ez(t) and Ey(t), are compared against respective thresholds, Tz and Ty, and a label L(t) is assigned to the current window data by the following classification rule: if Ey(t) > Ty, then L(t) = patient; else if Ez(t) > Tz, then L(t) = doctor; else, L(t) = pause. Note that the decision rule discards competing speech of provider and client.

A counter is used to track successively detected pauses. If the count exceeds a threshold Tc, then the current record of doctor's speech is ended and a new record is created. Speech records thus created captures from half a sentence up to five sentences, with majority of records containing one or two sentences. Subsequent speech acoustic model training and system evaluation tests are performed on speech records.

3.2. Acoustic model

Speech feature consisted of 39 components: 13 MFCC parameters $c_0, \cdots, c_{12}$ and their 1st and 2nd order time derivatives. Short-time analysis window size was 20 ms and shift was 10 ms. A total of 52 acoustic sound units were defined, including 42 speech monophone units, seven fill pause units, one unit for sound artifacts like lip smack and microphone ruffling, as well as one pause unit and one silence unit. Context-dependent triphone modeling was used for speech, and context independent modeling was used for the rest sound units. Gaussian mixture density based Hidden Markov model (HMM) with three emitting states was used for triphone and other sound units, where each emitting state was modeled by a size-16 Gaussian mixture density with diagonal covariance matrix. Acoustic model parameters were estimated by using the HTK toolkit [8], where HMM states were tied by phonetic decision trees.

3.3. Language model

The described captioning task involves spontaneous dialog style speech in various medical specialty domains. The major difficulty faced by the language modeling task here is that transcriptions of telehealth conversation were very limited, and there were little accessible textual corpora elsewhere that match the domain and style of telehealth. In order to improve quality of trained language model (LM), textual data from other domains were used to enlarge vocabulary coverage and improve events estimation.

Telehealth captioning requires good lexicon coverage of medical terms and effective estimation of small n-gram events containing medical terms. Toward this goal, words related to medicine were grouped into semantic classes, for example,

Table 1. Word classes defined for telehealth and examples

Description	Example
Disease	acalculia, paraplegia
Medicine	phenobarbital, precipitant
Human Body & Organs	follicle, capsule
Meditation Method	intubation, phototherapy
Medical Equipment & Facility	sigmoidoscopy, inhalator
Symptom(noun)	numb, stiffness
Symptom(adjective)	dizzy, drowsy
Medical &Chemical Object	peroxide, triglyceride
Profession Name	ophthalmologist, oncologist
Person Name	Garrett, Lewis
Number	One, two

names of medicines, diseases, therapeutic techniques, etc. In addition, digits and peoples' names were also categorized. The rest vocabulary words falling outside of these categories stand by themselves as singleton word classes. Table 1 shows the defined 11 classes. Performance of language model trained by the proposed word class definition was proven superior to either part of speech based or automatic clustering based word classes [9].

Supplementary text corpora used for LM training included public domain datasets of Switchboard, Broadcast News, and Call Home. In addition, a medical written report dataset was collected in this project to ensure good coverage of medical vocabulary, and an additional telehealth related dataset on dermatology was acquired. The datasets were categorized as in-domain and out-of-domain, where the in-domain ones included telehealth transcription sets and the telehealth related dermatology set, and the out-of-domain ones included Switchboard, Broadcast News, Call Home, and medical written reports.

Trigram LM was trained with Kneser-Ney backoff [10] by using the SRI toolkit [11]. Two ways of treating the in-domain and out-of-domain datasets were investigated. In the first case, class trigrams were trained for each dataset, generating six class trigram LMs. In the second case, two class trigram LMs were trained for the in-domain datasets, and four word trigrams LMs were trained for the out-of-domain datasets. In either case, the six LMs were linearly interpolated into a mixture LM by using a ten-fold validation on telehealth training set. The interpolation weights were optimized by a greedy forward selection algorithm [9] which combines a pair of LM at a time. The method yielded results slightly yet consistently better than commonly used EM algorithm.

The mixture LM of in-domain class trigrams and out-of-domain word trigrams, referred to as ICOW LM, was proven consistently the best over the mixture LM of all class trigram LMs as well as the mixture LM of all word trigram LMs. Significant performance gains were observed in reductions of test set perplexities which also translated into gains of recognition word accuracies for each speaker in speaker-dependent recognition tasks (Section 4).

3.4. Decoding engine

The speech decoding engine, TigerEngine v1.1, was developed in the Spoken Language and Information Processing Laboratory, Department of Computer Science of University of Missouri-Columbia, USA [6]. The decoding system performs large vocabulary continuous speech recognition in real-time based on one-pass time-synchronous Viterbi beam search, and its search organization is based on the lexical tree-copy algorithm [12]. Acoustic and language knowledge sources, including cross-word triphone HMMs and trigram LM are integrated as early as possible in search organization with a trigram LM lookahead.

An innovation feature of TigerEngine is a very fast and memory efficient language model lookup method for trigram-based language model lookahead, called Order-Preserving LM Context Pre-computing (OPCP). Specifically, OPCP efficiently builds a language model context array for each new LM context: minimum perfect hashing (MPH) is used to access the first LM score of the new context, and sequential access is used for the rest LM scores. The LM lookahead score for a node of a compressed lexical tree is obtained by maximizing over the trigram LM scores of a word list stored at the tree node. Fast LM access is attributed to the reduced number of hashing operations and the use of fast integer-key based hashing for the small number of hashing keys. Memory saving is achieved by storing only the last word index of trigram and by using MPH with small number of keys. OPCP reduced LM lookup time to about 10% total decoding time without decrease of word accuracy. The total memory cost of OPCP for LM lookup and storage was about the same or less than the original N-gram LM storage. Both the percentage of decoding time and the memory usage of OPCP are much less than commonly used methods in comparable decoders, and the time and memory advantages of OPCP become more pronounced with the increase of LM size.

3.5. Confidence annotation

Recognition outputs are further analyzed by a confidence annotation unit that utilizes novel features derived from confusion network (CN) and a random forest based classifier. A confusion network provides position-aligned competitive words and it is a linear graph transformed from word lattice. For real-time captioning, a fast confusion network generation algorithm was developed which took about 0.1xRT [13]. From CN, posterior entropy, posterior bigram and trigram LM scores are computed as confidence features, where posterior entropy measures the property of word posterior probabilities in the same position in a CN, and posterior bigram and trigram LM scores measure word n-gram probabilities conditioned on acoustic observation. Together with another novel feature of p-value that takes into account of Gaussian density spread better than likelihood in acoustic scores, and 8 previously proposed confidence features, a random forest based classifier is trained (see [14] for further details). The random forest based classifier achieved best performance in comparison with classifiers based on decision tree and support vector machine.

3.6. User interface

As was discussed in Section 1, captioning word rate needs to be constrained below 150 words / minute to be appropriate for slow readers. It is therefore desirable to implement a module that monitors word rate delivery by doctor and signals doctor to slow down when necessary. A side benefit is that by slowing down, doctor's speech may be better articulated and easier to be recognized. It was found that word rate estimated from recognition output word count aligned with speech input (2s window) was very close to the true word rate. A word rate signaling function based on such recognition word rate is being implemented into the system.

There is no doubt that caption errors need to be monitored to avoid causing confusions to patients. The above discussed automatic confidence annotation outputs are used to color code captions, for example, words with higher confidence scores are given darker shade than words with lower confidence scores. In

addition, the ultimate control on error correction is provided by the system to the doctor. Currently a pen based online caption correction system is under implementation where doctor can hold up conversation momentarily to check captions saved in word blocks. Errors that are deemed important by doctor can be corrected and the correct captions will be sent to patient.

4. EXPERIMENTAL RESULTS

The captioning task employed a vocabulary size of 46,489, with 3.07% of vocabulary word being medical terms. At the current time, speech recording data of five doctors, two females (D1 and D5) and three males (D2, D3, D4), were processed to train and test the system. A summary on the sizes of the datasets is provided in Table 2. The conversation dataset contains patients' speech, the training and test datasets together constitute doctor's speech data extracted from the conversation data, and a subset of each doctor's speech was set aside for use as test set. Word counts from transcription texts of the above doctors' speech data is also given in Table 2.

Table 2. Datasets of 5 doctors: speech (min.) /text (no. of words).

	Conversation	Training set	Test set
D1	630	210 / 35,348	29.8 / 5,105
D2	480	200 / 39,398	14.3 / 2,760
D3	300	145 / 28,700	19.3 / 3,238
D4	420	180 / 39,148	27.8 / 6,492
D5	380	250 / 44,967	21.1 / 3,998

Since doctors who use telehealth systems use them on regularly basis, current efforts on the captioning system are focused more on accurate captioning for individual doctors than on minimizing user enrollment time. Towards this end, five speaker dependent (SD) acoustic models were trained with one for each doctor, and one multi-speaker (MS) acoustic model was also trained as a reference. Similarly, five language models were trained for individual doctors and one for the five doctors. Test set perplexity of the five doctors D1 through D5 from the speaker-dependent ICOW LMs were 115.51, 84.49, 75.63, 116.84, and 107.12, respectively (lowest among all the LMs). Captioning word accuracies on test sets of the five doctors are shown in Table 3. It is observed that accuracy varies among doctors, ICOW gives best results for all five speakers, and SD models are much better than the MS model. Compared across speakers, test set perplexity did not correlate well with word accuracy. Rather, clarity of articulation, fluency of speech, speech rate were more critical factors in word accuracy. For example, D2 was recognized by listeners as a difficult speaker which correlated well with his low accuracy.

Table 3. Captioning word accuracy (%) on five doctors for three types of LMs. For ICOW LM, both SD and MS results are shown. Average accuracy was weighted by dataset sizes.

	Word-LM (SD)	Class-LM (SD)	ICOW-LM (SD/MS)
D1	79.45	80.14	80.81 / 77.53
D2	72.50	72.93	73.15 / 68.70
D3	73.63	74.05	73.81 / 71.68
D4	76.41	77.41	77.72 / 73.07
D5	80.67	80.99	81.34 / 78.31
Average	77.00	77.64	77.95 / 74.33

Confidence classification on speech recognition system outputs was produced by the random forest classifier with the 12 confidence features and a forest with 500 trees. The results are shown for the five doctors in Table 4, where for the reported error rate, the false alarm rate was at the level of approximately 3%.

Table 4. Confidence error rate (%) for the five doctors.

	D1	D2	D3	D4	D5	Average
C-Err	13.16	17.55	17.83	15.90	13.26	15.26

5. CONCLUSION

This paper presents a first exposition of an automatic captioning system designed for telemedicine system to assist patients with hearing loss in understanding doctor's questions and instructions. While the captioning system is primarily designed for hearing impaired clients, it is expected to be useful for normal hearing clients as well. The current system achieved a respectable performance for certain doctors participated in the study. Future work include collection of more data, improvement in acoustic and language model training, and evaluation and refinement of user interface.

REFERENCES

[1] http://www.cfv.org/caai/nadh22.htm

[2] http://www.NAD.org

[3] *Proc. DARPA Broadcast News Workshop,* Feb. 1999.

[4] A. Ando, T. Imai, A Kobayashi, H. Isono and K. Nakabayashi, "Real-time transcription system for simultaneous subtitling of Japanese broadcast news programs," *IEEE Trans. Broadcasting,* vol. 46, No. 3, pp. 189-196, September, 2000.

[5] M. Cettolo, F. Brugnara, and M. Federico, "Advances in the automatic transcription of lectures," *Proc. ICASSP,* pp. I-769-772, 2004.

[6] X. Li and Y. Zhao, "A fast and memory-efficient N-gram language model lookup method for large vocabulary continuous speech recognition," to appear in *Computer Speech & Language*, 2006.

[7] R. Hu, J. Xue and Y. Zhao, "Incremental largest margin linear regression and MAP adaptation for speech separation in Telehealth application," in *Proc. EuroSpeech*, pp. 261-264, Lisbon, Portugal, 2005.

[8] The HTK toolkit, http://htk.eng.cam.ac.uk/

[9] X. Zhang, "Language Modeling for Automatic Speech Recognition in Telehealth, MS thesis, Dept, of CS, Univ. of Missouri – Columbia, USA, Dec. 2005.

[10] R. Kneser R. and H. Ney, "Improved backing-off for M-gram language modeling," *in Proc. ICASSP,* pp. 181-184, 1995.

[11] A. Stolcke "SRILM - An extensible language modeling toolkit", in *Proc. ICSLP*, Denver, Colorado, September 2002.

[12] H. Ney & S. Ortmanns, "Dynamic programming search for continuous speech recognition," *IEEE Signal Processing Magazine* 16 (5): 64-83, 1999.

[13] J. Xue and Y. Zhao, "Improved confusion network algorithm and shortest path search from word lattice," in *Proc. ICASSP*, pp. I-854-857, Philadelphia, PA, March 2005.

[14]. J. Xue and Y. Zhao, "Random forest based confidence annotation using novel features from confusion network,"in *Proceedings of ICASSP*, May 2006.