Automated Speech Recognition Technology for Dialogue Interaction with Non-Native Interlocutors

Alexei V. Ivanov[†], Vikram Ramanarayanan[†], David Suendermann-Oeft[†], Melissa Lopez[‡], Keelan Evanini[‡] and Jidong Tao[‡]
Educational Testing Service R&D
[†] 90 New Montgomery St, # 1500, San Francisco, CA
[‡] 600 Rosedale Road, Princeton, NJ
{aivanou,vramanarayanan,suendermann-oeft,mlopez002,kevanini,jtao}@ets.org

Abstract

Dialogue interaction with remote interlocutors is a difficult application area for speech recognition technology because of the limited duration of acoustic context available for adaptation, the narrow-band and compressed signal encoding used in telecommunications, high variability of spontaneous speech and the processing time constraints. It is even more difficult in the case of interacting with non-native speakers because of the broader allophonic variation, less canonical prosodic patterns, a higher rate of false starts and incomplete words, unusual word choice and smaller probability to have a grammatically well formed sentence. We present a comparative study of various approaches to speech recognition in non-native context. Comparing systems in terms of their accuracy and real-time factor we find that a Kaldi-based Deep Neural Network Acoustic Model (DNN-AM) system with online speaker adaptation by far outperforms other available methods.

1 Introduction

Designing automatic speech recognition (ASR) and spoken language understanding (SLU) modules for spoken dialog systems (SDSs) poses more intricate challenges than standalone ASR systems, for many reasons. First, speech recognition latency is extremely important in a spoken dialog system for smooth operation and a good caller experience; one needs to ensure that recognition hypotheses are obtained in near real-time. Second, one needs to deal with the lack of (or minimal) context, since responses in dialogic situations can often be short and succinct. This also means that one might have to deal with minimal data for model adaptation. Third, these responses being typically spontaneous in nature, often exhibit pauses, hesitations and other disfluencies. Fourth, dialogic applications might have to deal with audio bandwidth limitations that will also have important implications for the recognizer design. For instance, in telephonic speech, the bandwidth (300-3200 Hz) is lesser than that of the hi-fidelity audio recorded at 44.1 kHz. All these issues can drive up the word error rate (WER) of the ASR component. In a recent study comparing several popular ASRs such as Kaldi (Povey et al., 2011), Pocketsphinx (Huggins-Daines et al., 2006) and cloud-based APIs from Apple[1], Google[2] and AT&T[3] in terms of their suitability for use in SDSs, In (Morbini et al., 2013) there was found no particular consensus on the best ASR, but observed that the open-source Kaldi ASR performed competently in comparison with the other closed-source industry-based APIs. Moreover, in a recent study, (Gaida et al., 2014) it was found that Kaldi significantly outperformed other open-source recognizers on recognition tasks on German Verbmobil and English Wall Street Journal corpora. The Kaldi online ASR was also shown to outperform the Google ASR API when integrated into the Czech-based ALEX spoken dialog framework (Plátek and Jurčíček, 2014).

The aforementioned issues with automatic speech recognition in SDSs are only exacerbated in the case of non-native speakers. Not only do non-native speakers pause, hesitate and make false starts more often than native speakers of a language, but their speech is also characterized by a broader allophonic variation, a less canonical prosodic pattern, a higher rate of incomplete words, unusual word choices and a lower probabil-

[1] Apple's Dictation is an OS level feature in both MacOSX and iOS.
[2] https://www.google.com/speech-api/v1/recognize
[3] https://service.research.att.com/smm

Proceedings of the SIGDIAL 2015 Conference, pages 134–138,
Prague, Czech Republic, 2-4 September 2015.

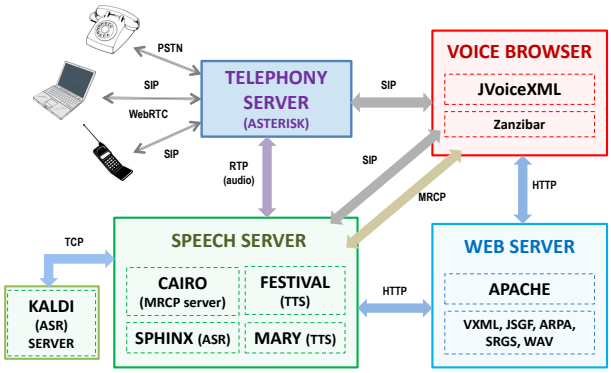

Figure 1: Architecture of the HALEF spoken dialog system.

ity of producing grammatically well-formed sentences. An important application scenario for non-native dialogic speech recognition is the case of conversation-based Computer-Assisted Language Learning (CALL) systems. For instance, *Subarashii* is an interactive dialog system for learning Japanese (Bernstein et al., 1999; Ehsani et al., 2000), where the ASR component of the system was built using the HTK speech recognizer (Young et al., 1993) with both native and non-native acoustic models. In general, the performance of the system after SLU was good for in-domain utterances, but not for out-of-domain utterances. As another example, in Robot Assisted Language Learning (Dong-Hoon and Chung, 2004) and CALL applications for Korean-speaking learners of English (Lee et al., 2010), whose authors showed that acoustic models trained on the Wall Street Journal corpus with an additional 17 hours of Korean children's transcribed English speech for adaptation produced as low as 22.8% WER across multiple domains tested. In the present work, we investigate the online and offline performance of a Kaldi Large Vocabulary Continuous Speech Recognition (LVCSR) system in conjunction with the open-source and distributed HALEF spoken dialog system (Mehrez et al., 2013; Suendermann-Oeft et al., 2015).

2 System description

Figure 1 schematically depicts the main components of the HALEF spoken dialog framework, of which the speech recognizer is a component. The various modules of HALEF include the Asterisk telephony server (van Meggelen et al., 2009), a voice browser based on JVoiceXML (Schnelle-Walka et al., 2013), a web server running Apache Tomcat, and a speech server, which consists of an MRCP server (Prylipko et al., 2011) in addition to text-to-speech (TTS) engines—Festival (Taylor et al., 1998) and Mary (Schröder and Trouvain, 2003)—as well as support for Sphinx-4 (Lamere et al., 2003) and Kaldi (Povey et al., 2011) ASRs. In contrast to Sphinx-4 which is tightly integrated into the speech server code base, Kaldi-based ASR is installed on an own server, which is communicating with the speech server via TCP socket. The advantages of this design decision are (a) the ease of management of the computational resources, required by Kaldi when operating in real-time mode (including the potential use of Graphical Processing Units (GPUs)), which could otherwise interfere with the other processes running on the speech server (audio streaming, TTS, Session Initiation Protocol (SIP) and Media Resource Control Protocol (MRCP) communication) and (b) the ease to test the very speech recognizer used in the live SDS also in the offline mode, for example for batch experiments. Often ASR configurations in live SDSs differ from batch systems that may result in different behaviour w.r.t. WER, latency, etc.

In this paper, we will be focusing specifically on evaluating the performance of the Kaldi ASR system within HALEF (we have already covered the Sphinx version in the papers cited above). We generally follow Kaldi's WSJ standard model generation recipe with a few modifications to accommodate our training data. The most sophisticated acoustic models are obtained with speaker adaptive training (SAT) on the feature Maximum Likelihood Linear Regression (fMLLR)-adapted data.

We use about 780 hours of non-native English speech to train the acoustic model. The speaker population covers a diversity of native languages, geographical locations and age groups. In order to match the audio quality standard of the Public Switched Telephone Network (PSTN), we reduce the sampling rate of our recordings down to 8 kHz. The language model was estimated on the manual transcriptions of the same training corpus consisting of $\approx$ 5.8 million tokens and finally was represented as a trigram language model with $\approx$ 525 thousand trigrams and $\approx$ 605 thousand bigrams over a lexicon of $\approx$ 23 thousand words which included entries for the most frequent partially produced words (e.g. ATTR-; ATTRA-; ATTRAC-

; ATTRACT; ATTRACT-; ATTRACTABLE). Ultimately, the final decoding graph was compiled having approximately 5.5 million states and 14 million arcs.

The default Kaldi speech recognizer use case is oriented towards optimal performance in transcription of large amounts of pre-recorded speech. In these circumstances there exists a possibility to perform several recognition passes and estimate the adaptation transformation from a substantial body of spoken material. The highest performing Deep Neural Network (DNN) acoustic model ("nnet2" in Kaldi notation) requires a prior processing pass with the highest performing Gaussian Mixture Model (GMM, "tri4b" in Kaldi notation), which in turn requires a prior processing pass with the same GMM in the speaker-independent mode.

However, in the dialogue environment, it is essential to be able to produce recognition results with the smallest possible latency and little adaptation material. That is the main reason for us to look for alternatives to the mentioned approach. One such possibility is to use the DNN acoustic model with un-adapted data and constrain its output via a speaker-dependent i-Vector (Dehak et al., 2011). This i-Vector contains information on centroids of the speaker-dependent GMM. The i-Vector can be continuously re-estimated based on the available up-to-the-moment acoustic evidence ("online" mode) or after presentation of the entire spoken content (the so called "offline" mode).

3 Experiments

The evaluation was performed using vocal productions obtained from language learners in the scope of large-scale internet-based language assessment. The production length is a major distinction of this data from the data one may expect to find in the spoken dialogue domain. The individual utterance is a quasi-spontaneous monologue elicited by a certain evaluation setup. The utterances were collected from six different test questions comprising two different speaking tasks: 1) providing an opinion based on personal experience and 2) summarizing or discussing material provided in a reading and/or listening passage. The longest utterances are expected to last up to a minute. The average speaking rate is about 2 words per second. Every speaker produces up to six such utterances. Speakers had a brief time to familiarize themselves with the task and prepare an approximate production plan. Although in strict terms, these productions are different from the true dialogue behavior, they are suitable for the purposes of the dialogic speech recognition system development.

The evaluation of the speech recognition system was performed using the data obtained in the same fashion as the training material. Two sets are used: the development set (dev), containing 593 utterances (68329 tokens, 3575 singletons, 0% OOV rate) coming from 100 speakers with the total amount of audio exceeding 9 hours; and the test set (test), that contains 599 utterances (68112 tokens, 3709 singletons, 0.18% OOV rate) coming from 100 speakers (also more than 9 hours of speech in total). We attempted to have a non-biased random speaker sampling, covering a broad range of native languages, English speaking proficiency levels, demographics, etc. However, no extensive effort has been spent to ensure that frequencies of the stratified sub-populations follow their natural distribution. Comparative results are presented in Table 1.

As it can be learned from Table 1, the "DNN i-Vector" method of speech recognition outperforms Kaldi's default "DNN fMLLR" setup. This can be explained by the higher variability of non-native speech. In this case the reduced complexity of the i-Vector speaker adaptation matches better the task that we attempt to solve. There is only a very minor degradation of the accuracy with the reduction of the i-Vector support data from the whole interaction to a single utterance. As expected, the "online" scenario loses some accuracy to the "offline" in the utterance beginning, as we could verify by analyzing multiple recognition results.

It is also important to notice that the accuracy of the "DNN i-Vector" system compares favorably with human performance in the same task. In fact, experts have the average WER of about 15% (Zechner, 2009), while Turkers in a crowdsourcing environment perform significantly worse, around 30% WER (Evanini et al., 2010). Our proposed system is therefore already approaching the level of broadly defined average human accuracy in the task of non-native speech transcription.

The "DNN i-Vector" ASR method vastly outperforms the baseline in terms of processing speed. Even with the large vocabulary model in a typical 10-second spoken turn we expect to have only 3 seconds of ASR-specific processing latency. Indeed, in order to obtain an expected de-

System	Adaptation	WER (dev)	WER (test)	xRT
GMM SI	Offline, whole interaction	37.58%	37.98%	0.46
GMM fMLLR	Offline, whole interaction	29.96%	31.73%	2.10
DNN fMLLR	Offline, whole interaction	22.58%	24.44%	3.44
DNN i-Vector	Online, whole interaction	21.87%	23.33%	1.11
DNN i-Vector	Offline, whole interaction	**21.81%**	23.29%	1.05
DNN i-Vector	Online, every utterance	22.01%	23.48%	**1.30**
DNN i-Vector	Offline, every utterance	21.90%	**23.22%**	1.13

Table 1: Accuracy and speed of the explored ASR configurations; WER – Word Error Rate; (dev) - as measured on the development set; (test) – as measured on the test set; xRT - Real Time factor, i.e. the ratio between processing time and audio duration; SI - Speaker Independent mode.

lay one shall subtract the duration of an utterance from the total processing time as the "online" recognizer commences speech processing at the moment that speech is started. That 3 seconds delay is very close to the natural inter-turn pause of 0.5 – 1.5 seconds. Better language modeling is expected to bring the xRT factor below one. The difference of the xRT factor between the "online" and "offline" modes can be explained with somewhat lower quality of acoustic normalization in the "online" case. Larger numbers of hypotheses fit within the decoder's search beam and, thus, increase the processing time.

4 Conclusions

The DNN i-Vector speech recognition method has proven to be sufficient in the task of supporting a dialogue interaction with non-native speakers. In respect to our baseline systems we observe improvements both in accuracy and processing speed. The "online" mode of operation appears particularly attractive because it allows to minimize the processing latency at the cost of a minor performance degradation. Indeed, the "online" recognizer is capable to start the processing simultaneously with the start of speech production. Thus, unlike the "offline" case, the total perceived latency in the case of "online" recognizer is xRT-1.

There are ways to improve our system by performing a more targeted language modeling and, possibly, language model adaptation to a specific dialogue turn. Our further efforts will be directed to reducing processing latency and increasing the system's robustness by incorporating interpretation feedback into the decoding process.

We plan to perform a comparative error analysis to have a better picture of how our automated system compares to the average human performance. It is important to separately evaluate WERs for the content vs functional word subgroups; determine the balance between insertions, deletions and substitutions in the optimal operating point; compare humans and machines in ability to recover back from the context of the mis-recognized word (e.g. a filler or false start).

We plan to collect actual spoken dialogue interactions to further refine our system through a crowdsourcing experiment in a language assessment task. Specifically, the ASR sub-sytem can benefit from sampling the elicited responses, measuring their apparent semantic uncertainty and tailoring system's lexicon and language model to better handle acoustic uncertainty of non-native speech.

References

Jared Bernstein, Amir Najmi, and Farzad Ehsani. 1999. Subarashii: Encounters in Japanese spoken language education. *CALICO Journal*, 16(3):361–384.

Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2011. Front–end factor analysis for speaker verication. In *IEEE Trans. on Audio, Speech and Language Processing*, volume 19, pages 788–798.

AHN Dong-Hoon and Minhwa Chung. 2004. One-pass semi-dynamic network decoding using a subnetwork caching model for large vocabulary continuous speech recongnition. *IEICE Trans. on Information and Systems*, 87(5):1164–1174.

Farzad Ehsani, Jared Bernstein, and Amir Najmi. 2000. An interactive dialog system for learning Japanese. *Speech Communication*, 30(2):167–177.

Keelan Evanini, Derrick Higgins, and Klaus Zechner. 2010. Using Amazon Mechanical Turk for

transcription of non-native speech. In *Proc. of the NAACL HLT Conference, Los Angeles, CA*.

Christian Gaida, Patrick Lange, Rico Petrick, Patrick Proba, Ahmed Malatawy, and David Suendermann-Oeft. 2014. Comparing open-source speech recognition toolkits. In *Technical Report, http://suendermann.com/su/pdf/oasis2014.pdf*.

D. Huggins-Daines, M. Kumar, A. Chan, A. Black, M. Ravishankar, and A. Rudnicky. 2006. Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France.

P. Lamere, P. Kwok, E. Gouvea, B. Raj, R. Singh, W. Walker, M. Warmuth, and P. Wolf. 2003. The CMU SPHINX-4 speech recognition system. In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, China.

Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, and G Lee. 2010. Postech approaches for dialog-based english conversation tutoring. *Proc. of Asia-Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference, Singapore*, pages 794–803.

T. Mehrez, A. Abdelkawy, Y. Heikal, P. Lange, H. Nabil, and D. Suendermann-Oeft. 2013. Who discovered the electron neutrino? A telephony-based distributed open-source standard-compliant spoken dialog system for question answering. In *Proc. of the German Society for Computational Linguistics (GSCL), Int. Conf. of the*, Darmstadt, Germany.

Fabrizio Morbini, Kartik Audhkhasi, Kenji Sagae, Ron Artstein, Dogan Can, Panayiotis Georgiou, Shri Narayanan, Anton Leuski, and David Traum. 2013. Which asr should i choose for my dialogue system. In *Proc. of the 14th annual SIGdial Meeting on Discourse and Dialogue, Metz, France*, pages 394–403.

Ondřej Plátek and Filip Jurčíček. 2014. Integration of an on-line kaldi speech recogniser to the Alex dialogue systems framework. In *Text, Speech and Dialogue*, pages 603–610. Springer.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi speech recognition toolkit. In *Proc. of the Automatic Speech Recognition and Understanding (ASRU), Int. Workshop on*, Hawaii, USA.

D. Prylipko, D. Schnelle-Walka, S. Lord, and A. Wendemuth. 2011. Zanzibar OpenIVR: an open-source framework for development of spoken dialog systems. In *Proc. of the Text, Speech and Dialogue (TSD), Int. Conf. on*, Pilsen, Czech Republic.

D. Schnelle-Walka, S. Radomski, and M. Mühlhäuser. 2013. JVoiceXML as a modality component in the W3C multimodal architecture. *Journal on Multimodal User Interfaces*, 7:183–194.

Marc Schröder and Jürgen Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

David Suendermann-Oeft, Vikram Ramanarayanan, Moritz Teckenbrock, Felix Neutatz, and Dennis Schmidt. 2015. Halef: an open-source standard-compliant telephony-based modular spoken dialog system–a review and an outlook. In *International Workshop on Spoken Dialog Systems (IWSDS) 2015, Busan, South Korea*.

P. Taylor, A. Black, and R. Caley. 1998. The architecture of the Festival speech synthesis system. In *Proc. of the ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

J. van Meggelen, J. Smith, and L. Madsen. 2009. *Asterisk: The Future of Telephony*. O'Reilly, Sebastopol, USA.

S. Young, P. Woodland, and W. Byrne. 1993. *The HTK Book, Version 1.5*. Cambridge University, Cambridge, UK.

Klaus Zechner. 2009. What did they actually say? Agreement and disagreement among transcribers of non-native spontaneous speech responses in an English proficiency test. In *Proc. of the ISCA SLaTE Workshop, Wroxall Abbey Estate, Warwickshire, England*.