Medical Speech Recognition: Reaching Parity with Humans

Erik Edwards, Wael Salloum, Greg P. Finley, James Fone,
Greg Cardiff, Mark Miller, and David Suendermann-Oeft

EMR.AI Inc., San Francisco, CA, USA
david@emr.ai,
http://emr.ai

Abstract. We present a speech recognition system for the medical domain whose architecture is based on a state-of-the-art stack trained on over 270 hours of medical speech data and 30 million tokens of text from clinical episodes. Despite the acoustic challenges and linguistic complexity of the domain, we were able to reduce the system's word error rate to below 16% in a realistic clinical use case. To further benchmark our system, we determined the human word error rate on a corpus covering a wide variety of speakers, working with multiple medical transcriptionists, and found that our speech recognition system performs on a par with humans.

Keywords: medical speech recognition, human word error rate, parity

1 Introduction

There are several unique challenges in medical-domain automatic speech recognition (ASR). Acoustically, there is often significant background noise during medical dictations, for example, ranging from sirens to office noises to competing talkers and background conversations. This is a challenge not only because of the lower signal-to-noise ratio, but also because of the diversity of acoustic environments and noise sources which may affect a given recording. This is different than, say, a system designed to work with car noise or other known noise sources. Second, a large number of microphones and dictation devices are used, e.g. Dictaphone, SpeechMike, Digital Voice Recorders, or regular telephone handsets, many of which use special lossy codecs of varying sound quality. Also, the speakers do not maintain any consistent distance or orientation to the microphone, such that even a single utterance can vary widely in quality and absolute volume.

Speech in medical dictations differs from normal (conversational) speech in several ways, although the presence and magnitude of these differences varies from one recording to another. Thus, dictations are often spoken very rapidly, but other files are found with slower speech and lengthy hesitations. Many voiced hesitations are also present, as are repeated words and restarted sentences. Sentences often lack clear juncture, boundaries, or formatting commands. Even itemized lists are sometimes spoken in rapid succession that is unrevealing of boundaries. One often gets the impression that the speech is spoken without a sense of a human listener, or even the intention of being understood, but rather only for required cataloging purposes. On the other hand, some dictations are spoken with excellent quality by certain physicians or their professional assistants.

Finally, and perhaps most importantly, is the challenge of highly complex, domain-specific medical terminology, including thousands of drug names. This presents a significant out-of-vocabulary challenge and is perhaps why most existing medical ASR work has only taken on a single domain (usually radiology).

Although medical-domain ASR has been reported in some form since the 1980s [1–3], all work prior to 1999 used single-word as opposed to continuous ASR, with a single early exception for German [4]. Early works on continuous medical ASR [5–7] immediately recognized the importance of including medical domain-specific terminology in the statistical language model. However, the physicians (usually radiologists) were themselves enlisted to provide manual corrections to update the ASR lexicon. Only gradually in the 21st century have a handful of studies begun to use non-physician transcriptions for language model training. Most reports come from the single domain of radiology (e.g. [8, 9]), although a smaller number of restricted-domain systems have been reported elsewhere (dermatology: [10]; temporomandibular disorder: [11]). We are developing language model methodology that scales to larger volumes of data from multiple subspecialties.

We found 45 studies since 1999 that assess the quality of medical ASR, e.g. those covered in reviews by Hodgson et al. [12] and Hammana et al. [13]. Among the few publications on speech recognition on medical corpora reporting results in terms of Word Error Rate (WER) is the work by [14] and [15] on clinical question answering. The latter focuses on spontaneously spoken medical questions and reports 29.3% for the SRI Decipher system and 37.3% for Nuance Dragon, Medical version. Both systems were adapted to the specific study domain by language model adaptation. Nuance Dragon also underwent profile training to enhance performance. Luu et al. [16] covers nursing handover and reports a WER of 24.6%. Paats et al. [17] and Alumäe [18] both cover radiology reports in Estonian and report WERs of 18.4% and 13.7%, respectively.

We found ten studies since 1999 that compare ASR quality with human transcription (HT) in healthcare. All 10 report more errors with ASR than HT, often substantially more [12, 19–21]. Where categorized, serious errors were also greater with ASR than HT. For example du Toit et al. [22] report 9.6% of ASR'd and 2.3% of HT'd charts having 'clinically significant' errors, and Basma et al. [23] conclude that 'major' errors were 8 times more likely with ASR than HT.

We report here on our current progress in developing a medical ASR system whose initial version was discussed in [24]. As indicated in the abstract, our system approaches human WER in the medical transcription domain covered by the corpus described in Section 2. Details on how we determined human transcription performance on this corpus are provided in Section 3. Reaching performance parity in this work was not due to novelty in one particular part of the ASR methodology, but rather to the accumulation of advances in all stages of the ASR training and decoding. Therefore, this paper presents an overview of our system with commentary on each of the stages in Section 4. Results are being discussed in Section 5 followed by conclusion and future outlook in Section 6.

2 Corpus of Medical Dictation

The studies described in this paper were carried out using a massive collection of English dictated out-patient reports covering a variety of different medical specializations. To perform the experiments whose details are provided in the sections following, three different types of corpora were required:

– The first corpus, (**M1**), contains both audio recordings and textual transcriptions of a small number of prototypical speakers covering the whole spectrum of difficulty levels of clinical dictation. We selected a total of nine speakers reaching from excellent, almost professional speakers dictating in clean office conditions, providing grammatically accurate sentences and punctuation commands, all the way to ones who dictated in an extreme rush, mumbling with no natural pauses, flat intonation, and extreme background noise and reverberation.

		M1	M2	M30
Train	Episodes	1 818	4 574	33 684
	Speakers	9	233	
	Duration/h	67	204	
	Tokens	620 926	1 629 469	29 846 087
	Types	9 872	20 203	68 369
	Singletons	3 451 (35%)	5 968 (30%)	15 613 (23%)
Test	Episodes	30	88	500
	Speakers	6	60	
	Duration/h	1.0	3.7	
	Tokens	10 077	28 696	138 792

Table 1. Corpus statistics

- The second corpus, (**M2**), features a random sample of audio and transcriptions of over 200 speakers in a hospital network representing the natural distribution of users in an operational scenario.
- The third corpus, (**M30**), consists of dictations of over 30 thousand outpatient letters used to build the language model for the clinical speech recognizer.

Detailed statistics of these corpora are provided in Table 1.

3 Human Baseline Performance

Multiple methods to determine human baseline performance on transcription tasks of differing complexity have been discussed in literature. [25] had several expert transcribers transcribe long spontaneous utterances of English language learners and compared their transcriptions with respect to word error rate. In a second phase, transcribers were allowed to choose preferred transcriptions from the set of available ones and correct them, and in a third phase, a gold standard transcription was picked by majority vote among all transcribers. Human word error rates varied between 20.5% in Phase 1 and 5.1% in Phase 3. More recently, [26] used a two-pass transcription approach comparing a first draft transcription with a second pass corrected version of another listener. Resulting human word error rates were reported on two standard research corpora, 5.9% on Switchboard and 11.3% on CallHome. [27] also measured the human error rate on these standard corpora by coupling three junior transcribers with a senior listener who performed error correction. The best resulting error rates were 5.1% on Switchboard and 6.8% on CallHome. Especially the discrepancy of the reported results on CallHome indicates that the measurement of human baseline performance on speech recognition is not a straightforward task.

In the present work, we followed a similar approach to these recent publications in that we compared a single pass transcription with one that went through multiple rounds of quality improvements. These rounds included: First draft of the medical report, quality assurance of the report to the level that it could be delivered as out-patient letter, and, finally, assuring that transcription guidelines were properly followed, e.g. that every uttered word is spelled out. The transcribers used in this study were professionally trained medical transcriptionists embedded in a private crowd [28], as described in further detail in [29].

In order to cover the full spectrum of difficulty levels of human and automatic transcription, we chose the M1 test set for this study. It contains a variety of different recording and speaking conditions, and has a size of over ten thousand tokens to reliably test for statistical significance of performance differences. The human word error rate achieved for this set following the above described methodology was 17.4%.

4 Acoustic and Language Model Training

At the time of decoding, the ASR system requires a language model (LM) and an acoustic model (AM). The former represents N-gram statistics of words, obtained from text processing. The latter represents a mapping from an audio file to phonology, which provides the link to the LM via the lexicon. The lexicon is an a priori mapping from words to phonological representation (pronunciations). The AM training proceeds in three global stages (Figure 1): a) feature extraction; b) alignments; c) DNN training. During decoding, only a) and c) are used; i.e., the features are extracted from test data, and the final DNN model provides the nonlinear mapping to phonology. The linguistic probabilities are represented by finite-state transducers (FSTs) [30, 31], which are implemented e.g. in Sphinx-4 [32], the OpenFST library [33, 34], and Kaldi [35].

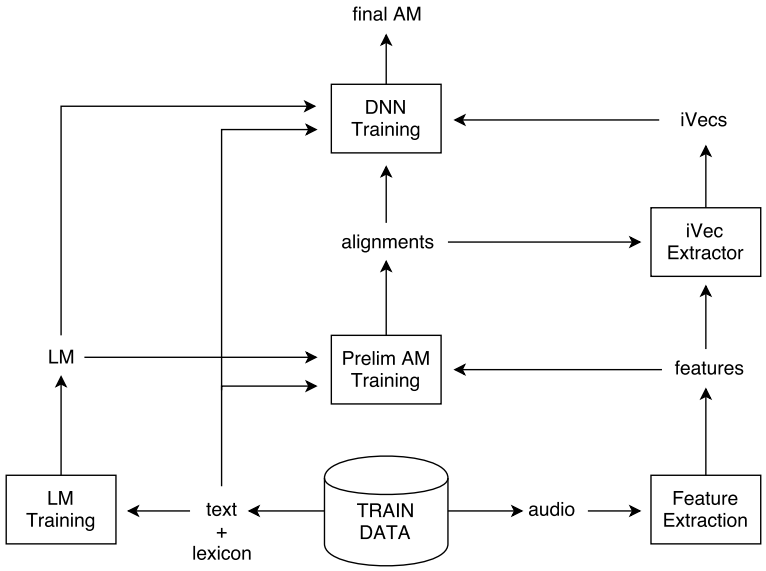

Fig. 1. Overview of acoustic model (AM) training for ASR. The deep neural network (DNN) training takes LM input, alignments (between phonemes and features time-line), and iVectors (iVecs) that are extracted from the audio features.

a) Feature extraction. This stage transforms the raw acoustic waveform, $w(n)$, sampled at 8-48 kHz, into a multivariate time-series of C features, sampled at 100 Hz frame rate. As it turns out, the raw waveform in medical dictations presents with widely-varying dynamic range. Some recording systems use spectral subtraction [36] or other speech enhancement [37], which can result in sections of near-absolute silence in some audio files; other sections include strong background noises (e.g. sirens) and loud speech; and a typical medical-dictation audio file includes more variability in speech volume and orientation w.r.t. the microphone than typical in standard speech corpora. Therefore, we have carefully considered waveform root-mean-square (RMS) normalization, power-normalization of the spectrogram, and mean/variance normalization of the final features. First, contiguous sections of near-silence (below 0.1% of max level) are clipped out (Figure 2). Second, the RMS (squared signal smoothed with 200-ms time-constant) is normalized such that all audio files are scaled to a common 70 dB sound-pressure level in units of pressure (Pascals), where 70 dB is chosen as the typical normal-to-

loud conversational range [38]. Whereas perceptual loudness correlates to recent amplitude maxima [39], the use of maxima for amplitude normalization is unstable, as extrema are always subject to greater statistical variation. We chose the RMS 90-percentile for each audio file as the normalization point, which we found to be more homogeneous across files. Third, we considered three methods of power-normalization following the short-term Fourier transform (STFT) log compression (standard), the nonlinear power-normalization (PN) method of Kim and Stern [40, 41], and their simple PN (sPN) method, which is first-order running-mean normalization. Preliminary results indicate PN as the preferred method, but these studies are ongoing and all results reported here use traditional log compression. Finally, the feature time-series are subjected to per-utterance mean and variance normalization (z-score method of Figure 2), or per-speaker mean and variance normalization, as in [35]. We are currently exploring the options shown in Figure 2 with very promising results, but all results presented in this paper use traditional mel-frequency cepstral coefficients (MFCCs) [42] utilizing the following options: 25-ms Hamming window, mel-frequency scale [43], no spectral transformations by PLP [44] or MVDR [45, 46], log compression, cepstral coefficients (CCs), and the typical lifter (cepstral-domain weighting) used in speech processing (as given by Juang et al. [47]). For DNN training, we use 40-dim MFCCs, as suggested by the study of Rath et al. [48], but for alignments we use 13-dim MFCCs with deltas and delta-deltas.

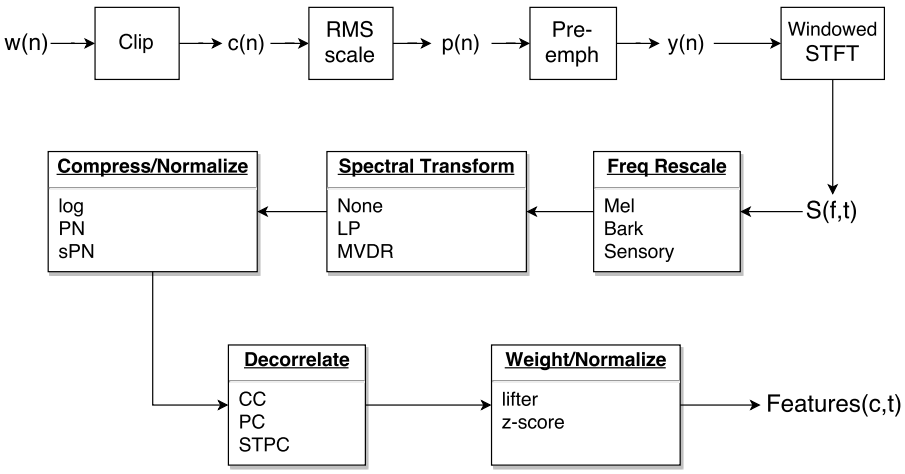

Fig. 2. Overview of acoustic feature extraction for a single utterance. The raw sampled waveform, $w(n)$, corresponding to a single wav file, is clipped of silent sections, and then RMS scaled into units of pressure in Pascals, $p(n)$. This is subjected to standard 1st-order pre-emphasis, before the windowed STFT. Note that n indexes samples at 8 to 48 kHz, whereas t indexes frames at 100 Hz. The subsequent stages are listed with options as described in the text.

b) Alignments. Transcribed medical dictations provide the correct word sequences for training, but no temporal information. Alignment to the audio file requires learning a model to map from acoustic features to phonological sequences, along with the lexicon to map from phonology to words. Several toolboxes can be used to obtain alignments, for example Praat [49], HTK [50], Julius [51], Kaldi [35], or RASR [52]. We do not conceive of alignment as a generic pipeline which is run on the data set at large. Rather, a best alignment can be obtained for each utterance and retained in the database, with each potentially derived from independent sources. The results reported here are based on triphone models [53], implementing a pipeline of speaker-independent Gaussian mixture models (GMMs) and linear discriminant

analysis + maximum-likelihood linear transformation (LDA+MLLT), followed by speaker-adaptive training (SAT) using LDA+MLLT and feature-space maximum-likelihood linear regression (fMLLR).

c) Deep neural network (DNN) training. Although artificial neural networks have been attempted since the late 1980s in ASR [54], and steadily advanced over the ensuing decades [55, 56], they have only become the most widely used state-of-the-art method in ASR in the last five years, joining other fields in the deep learning revolution. For example, the initial publication of the Kaldi toolkit [35] does not mention DNNs, and the recent theses of Plátek [57] and Gil [58] use Kaldi for ASR, but no DNNs. Kaldi introduced two DNN methods circa 2013 [48, 59], which we explored along with several other general machine learning toolboxes (Theano, etc.) for DNN training. For example, Miao [60] developed a hybrid Kaldi-Theano ASR system.

Another important part of current state-of-the-art ASR practice is the use of i-Vectors (iVecs) for training the DNN (Figure 1). These are derived by passing the features through a GMM-based universal background model (UBM), previously trained on the whole corpus [61, 62]. iVecs were introduced in 2009 for speaker recognition [63], brought into ASR work in 2011 [64, 65], and just recently used with DNNs for ASR model training [66, 67, 62]. We specify these dates to reinforce our general point that: although medical ASR was somewhat negatively viewed one decade ago, it is understandable that newer reports and reviews become increasingly optimistic. The field of medical ASR is likely only at the beginning of the change-over to DNN methods and the possibilities implied by near-human performance levels.

The language model used in this work is a conventional trigram model with Kneser-Ney smoothing. Discounting parameters were optimized by minimizing perplexity on a held-out set from the M1 training set. Training data comprised both the manual transcriptions in M1 and M2, and the outpatient letters in M30. The latter differ from transcriptions in that they contain case distinctions, formatting, punctuations, and numerals that are either absent or spelled out in transcriptions—e.g., a spoken 'colon' in transcriptions versus ':' in letters, 'twenty-three' in transcriptions versus '23' in letters. Prior to LM training, we processed M30 to remove formatting and spell out those characters and numerals that are typically spelled out in transcriptions.

Investigations into more sophisticated language modeling techniques are currently carried out, examples of which are given in Section 6. They will be subject to a future review publication.

5 Experimental Results and Discussion

As indicated in Section 2, in this study, we carried out two major experiments. The first one was dedicated to comparing human transcription performance on medical dictation to the performance of our speech recognition system on a range of difficulty levels. The second one was to investigate how the presented speech recognition system performs on a comprehensive selection of speakers, following the distribution in a realistic clinical use case. In the following, we will present and discuss results of these two experiments.

5.1 Comparing Human and ASR Performance

We trained the recognition models according to the pipeline described in Section 4 using all available speech data (M1 and M2 training sets) for the acoustic model and the transcriptions of the very same data for the language model. For evaluation, we used the M1 test set as motivated in Section 2. Table 2 shows the results of this experiment and compares them to the human baseline performance established in Section 3. The achieved performance of our speech

	Errors	Tokens	WER
ASR	1 850	10 115	18.3%
Human Baseline	1 760	10 115	17.4%

Table 2. Comparing Human and ASR Performance

Errors	Tokens	WER
4 413	28 696	15.4%

Table 3. ASR Performance in a Realistic Clinical Use Case

recognizer in this task was 18.3% WER which is less than one percentage point higher than the human baseline. To test whether this performance difference is of statistical significance, we carried out a two-proportion z-test. The resulting p-value is 0.10 which suggests that the difference observed in this experiment was *not* statistically significant at the $p < 0.05$ level, despite the rather large test set comprising over ten thousand tokens. While increasing the size of the test set will eventually reveal which of the two, human or machine, outperforms the other, this experiment shows that the accuracy of the speech recognizer we constructed is only marginally different from that of a professional medical transcriptionist, and is, hence, reaching parity.

5.2 ASR Performance in a Realistic Clinical Use Case

For the second experiment, we used the same acoustic model as before, trained on the M1 and M2 training sets, i.e. a total of 271 hours of speech. In order to prepare for a deployment in a realistic clinical use case, we substantially increased the size of the language model training corpus by including another 30 million tokens of medical reports (the training set of M30). This time, the experiment was carried out on the M2 test set which matches the target distribution in a realistic clinical setting. The results are shown in Table 3. The error rate, 15.4% is statistically significantly lower than that reported in the first experiment and establishes a strong baseline performance for a realistic clinical use case.

6 Conclusion and Future Directions

We have shown that a carefully tuned state-of-the-art speech recognizer, whose acoustic and language models were trained on moderate size speech and language corpora covering speech and relevant report samples of a set of over 270 physicians, is able to perform on a par with professional human medical transcribers. The human performance was measured in a single pass scenario, i.e., with no additional quality assurance or automated assistance (apart from a spell checker). Following previous work on measuring and optimizing human performance, e.g., in multi-pass or quality control scenarios, indicates that the human word error rate can be further improved. However, as it stands, the presented speech recognizer could be capable of serving as an automated first-pass transcriptionist. Furthermore, the authors are currently working on a number of enhancements to the speech recognizer which should result in substantial further improvements to the error rate, including

– optimizing the feature extraction configuration—the graph in Figure 2 shows our feature extraction pipeline and the diverse algorithms which we can choose from

– optimizing speaker clustering—we have seen significant performance gains by splitting speakers into specific speaker groups by certain criteria (e.g. region, gender, native language); our goal is to find the optimal split to optimize overall word error rate
– unsupervised acoustic model adaptation—making use of tens of thousands of hours untranscribed speech
– enhancing the language model by a) adding substantially more data (several million episodes), b) using sophisticated interpolation techniques, and c) rescoring with RNN-based, skip, or class language models.

References

1. Leeming, B., Porter, D., Jackson, J., Bleich, H., Simon, M.: Computerized radiologic reporting with voice data-entry. Radiology **138**(3) (1981) 585–588
2. Akers, G.: Using your voice: speech recognition technology in medicine and surgery. Clin Plast Surg **13**(3) (1986) 509–511
3. Matumoto, T., Iinuma, T., Tateno, Y., Ikehira, H., Yamasaki, Y., Fukuhisa, K., Tsunemoto, H., Shishido, F., Kubo, Y., Inamura, K.: Automatic radiologic reporting system using speech recognition. Med Prog Technol **12**(3-4) (1987) 243–257
4. Steinbiss, V., Ney, H., Essen, U., Tran, B.H., Aubert, X., Dugast, C., Kneser, R., Meier, H.G., Oerder, M., Haeb-Umbach, R., Geller, D., Höllerbauer, W., Bartosik, H.: Continuous speech dictation from theory to practice. Speech Commun **17**(1-2) (1995) 19–38
5. Hundt, W., Stark, O., Scharnberg, B., Hold, M., Kohz, P., Lienemann, A., Bonél, H., Reiser, M.: Speech processing in radiology. Eur Radiol **9**(7) (1999) 1451–1456
6. Zafar, A., Overhage, J., McDonald, C.: Continuous speech recognition for clinicians. J Am Med Inform Assoc **6**(3) (1999) 195–204
7. Devine, E., Gaehde, S., Curtis, A.: Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. J Am Med Inform Assoc **7**(5) (2000) 462–468
8. Paulett, J., Langlotz, C.: Improving language models for radiology speech recognition. J Biomed Inform **42**(1) (2009) 53–58
9. Hawkins, C., Hall, S., Hardin, J., Salisbury, S., Towbin, A.: Prepopulated radiology report templates: a prospective analysis of error rate and turnaround time. J Digit Imaging **25**(4) (2012) 504–511
10. Smith, K.: A discrete speech recognition system for dermatology: 8 years of daily experience in a medical dermatology office. Semin Cutan Med Surg **21**(3) (2002) 205–208
11. Hippmann, R., Dostálová, T., Zvárová, J., Nagy, M., Seydlová, M., Hanzlícek, P., Kriz, P., mídl, L., Trmal, J.: Voice-supported electronic health record for temporomandibular joint disorders. Methods Inf Med **49**(2) (2010) 168–172
12. Hodgson, T., Coiera, E.: Risks and benefits of speech recognition for clinical documentation: a systematic review. J Am Med Inform Assoc **23**(e1) (2016) e169–e179
13. Hammana, I., Lepanto, L., Poder, T., Bellemare, C., Ly, M.S.: Speech recognition in the radiology department: a systematic review. HIM J **44**(2) (2015) 4–10
14. Cao, Y.g., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J., Ely, J., Yu, H.: Askhermes: an online question answering system for complex clinical questions. J Biomed Inform **44**(2) (2011) 277–288
15. Liu, F., Tur, G., Hakkani-Tür, D., Yu, H.: Towards spoken clinical-question answering: evaluating and adapting automatic speech-recognition systems for spoken clinical questions. J Am Med Inform Assoc **18**(5) (2011) 625–630
16. Luu, T., Phan, R., Davey, R., Hanlen, L., Chetty, G.: Automatic clinical speech recognition for clef 2015 ehealth challenge. Working notes report/paper, University of Canberra (2015)
17. Paats, A., Alumäe, T., Meister, E., Fridolin, I.: Evaluation of automatic speech recognition prototype for estonian language in radiology domain: a pilot study. In: Proc Nordic-Baltic Conf Biomed Eng, Gothenburg, Sweden, Springer (2015) 96–99

18. Alumäe, T.: Full-duplex speech-to-text system for estonian. In: Proc Baltic HLT, Kaunas, Lithuania, IOS Press (2014) 3–10
19. du Toit, J., Hattingh, R., Pitcher, R.: The accuracy of radiology speech recognition reports in a multilingual south african teaching hospital. BMC Med Imaging **15**(8) (2015) 1–
20. Strahan, R., Schneider-Kolsky, M.: Voice recognition versus transcriptionist: error rates and productivity in mri reporting. J Med Imaging Radiat Oncol **54**(5) (2010) 411–414
21. Zick, R., Olsen, J.: Voice recognition software versus a traditional transcription service for physician charting in the ed. Am J Emerg Med **19**(4) (2001) 295–298
22. DuToit, J., Hattingh, R., Pitcher, R.: The accuracy of radiology speech recognition reports in a multilingual South African teaching hospital. BMC Medical Imaging **15**(8) (2015)
23. Basma, S., Lord, B., Jacks, L., Rizk, M., Scaranelo, A.: Error rates in breast imaging reports: comparison of automatic speech recognition and dictation transcription. AJR Am J Roentgenol **197**(4) (2011) 923–927
24. Suendermann-Oeft, D., Ghaffarzadegan, S., Edwards, E., Salloum, W., Miller, M.: A system for automated extraction of clinical standard codes in spoken medical reports. In: Proc Wrkshp SLT, San Diego, CA, IEEE (2016)
25. Zechner, K.: What did they actually say? agreement and disagreement among transcribers of non-native spontaneous speech responses in an english proficiency test. In: Proc SLaTE, Warwickshire, UK, ISCA (2009) 25–28
26. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G.: Achieving human parity in conversational speech recognition. arXiv **1610**(05256) (2017) 1–13
27. Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.L., Roomi, B., Hall, P.: English conversational telephone speech recognition by humans and machines. arXiv **1703**(02136) (2017) 1–7
28. Suendermann, D., Pieraccini, R.: Crowdsourcing for industrial spoken dialog systems. In Eskénazi, M., Levow, G.A., Meng, H., Parent, G., Suendermann, D., eds.: Crowdsourcing for speech processing. J. Wiley, Chichester (2013) 280–302
29. Salloum, W., Edwards, E., Ghaffarzadegan, S., Suendermann-Oeft, D., Miller, M.: Crowdsourced continuous improvement of medical speech recognition. In: Proc AAAI Wrkshp Crowdsourcing, San Francisco, CA, AAAI (2017)
30. Glass, J., Hazen, T., Hetherington, I.: Real-time telephone-based speech recognition in the jupiter domain. In: Proc ICASSP. Volume 1., IEEE (1999) 61–64
31. Mohri, M., Pereira, F., Riley, M.: Weighted finite-state transducers in speech recognition. Comput Speech Lang **16**(1) (2002) 69–88
32. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvêa, E., Wolf, P., Woelfel, J.: Sphinx-4: a flexible open source framework for speech recognition. Technical Report SMLI TR2004-0811, Sun Microsystems, Inc. (2004)
33. Allauzen, C., Riley, M., Schalkwyk, J., Mohri, M.: Openfst: a general and efficient weighted finite-state transducer library. Lect Notes Comput Sci **4783** (2007) 11–23
34. Gorman, K.: Openfst library: http://openfst.org (2016)
35. Povey, D., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlícek, P., Qian, Y., Schwarz, P., Silovsky, J.: The kaldi speech recognition toolkit. In: Proc Wrkshp ASRU, IEEE (2011) 4 p.
36. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust **27**(2) (1979) 113–120
37. Hermus, K., Wambacq, P., Van hamme, H.: A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP J Adv Signal Process **2007**(45821) (2007) 1–15
38. Zwicker, E., Feldtkeller, R.: Das Ohr als Nachrichtenempfänger. 2nd ed. edn. Monographien der elektrischen Nachrichtentechnik ; Bd. 19. Hirzel, Stuttgart (1967)
39. Fastl, H., Zwicker, E.: Psychoacoustics: facts and models. 3rd ed. edn. Springer, Berlin; New York (2007)
40. Kim, C., Stern, R.: Power-normalized cepstral coefficients (pncc) for robust speech recognition. In: Proc ICASSP, IEEE (2012) 4101–4104
41. Kim, C., Stern, R.: Power-normalized cepstral coefficients (pncc) for robust speech recognition. IEEE/ACM Trans Audio Speech Lang Processing **24**(7) (2016) 1315–1329

42. Imai, S.: Cepstral analysis synthesis on the mel frequency scale. In: Proc ICASSP. Volume 8., IEEE (1983) 93–96
43. Stevens, S., Volkmann, J.: The relation of pitch to frequency: a revised scale. Am J Psychol **53**(3) (1940) 329–353
44. Hermansky, H.: An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory perception. In: Proc ICASSP. Volume 12., IEEE (1987) 1159–1162
45. Murthi, M., Rao, B.: Minimum variance distortionless response (mvdr) modeling of voiced speech. In: Proc ICASSP. Volume 3., IEEE (1997) 1687–1690
46. Yapanel, U., Dharanipragada, S., Hansen, J.: Perceptual mvdr-based cepstral coefficients (pmccs) for high accuracy speech recognition. In: Proc EUROSPEECH, ISCA (2003) 1829–1832
47. Juang, B.H., Rabiner, L., Wilpon, J.: On the use of bandpass liftering in speech recognition. IEEE Trans Acoust **35**(7) (1987) 947–954
48. Rath, S., Povey, D., Veselý, K., Cernocký, J.: Improved feature processing for deep neural networks. In: Proc INTERSPEECH, ISCA (2013) 109–113
49. Boersma, P., van Heuven, V.: Praat, a system for doing phonetics by computer. Glot Int **5**(9/10) (2002) 341–345
50. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The htk book (for htk version 3.4). Book HTK Version 3.4, Cambridge Univ. Engineering Dept. (3 2009)
51. Lee, A.: The julius book. Book, Nagoya Institute of Technology (5 2010)
52. Rybach, D., Hahn, S., Lehnen, P., Nolden, D., Sundermeyer, M., Tüske, Z., Wiesler, S., Schlüter, R., Ney, H.: Rasr the rwth aachen university open source speech recognition toolkit. In: Proc Wrkshp ASRU, IEEE (2011) 4 p.
53. Gaida, C., Lange, P., Petrick, R., Proba, P., Malatawy, A., Suendermann-Oeft, D.: Comparing open-source speech recognition toolkits. Technical report, DHBW (10 2014)
54. Lippmann, R.: Review of neural networks for speech recognition. Neural Comput **1**(1) (1989) 1–38
55. Bourlard, H., Morgan, N., Renals, S.: Neural nets and hidden markov models: review and generalizations. Speech Commun **11**(2-3) (1992) 237–246
56. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process Mag **29**(6) (2006) 82–97
57. Plátek, O.: Speech recognition using Kaldi. Masters thesis, Charles Univ. (2014)
58. Gil, V.: Automatic speech recognition with Kaldi toolkit. Doctoral thesis, Univ. Politècnica de Catalunya (2016)
59. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.: Improving deep neural network acoustic models using generalized maxout networks. In: Proc ICASSP, IEEE (2014) 215–219
60. Miao, Y.: Kaldi + pdnn: building dnn-based asr systems with kaldi and pdnn. arXiv **1401.6984** (2014) 4 p.
61. Povey, D., Chu, S., Varadarajan, B.: Universal background model based speech recognition. In: Proc ICASSP, IEEE (2008) 4561–4564
62. Snyder, D., Garcia-Romero, D., Povey, D.: Time delay deep neural network-based universal background models for speaker recognition. In: Proc Wrkshp ASRU, IEEE (2015) 92–97
63. Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P.: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proc INTERSPEECH, ISCA (2009) 1559–1562
64. Zhang, Y., Yan, Z.J., Huo, Q.: A new i-vector approach and its application to irrelevant variability normalization based acoustic model training. In: Proc Wrkshp MLSP, IEEE (2011) 1–6
65. Karafiát, M., Burget, L., Matejka, P., Glembek, O., Cernocký, J.: ivector-based discriminative adaptation for automatic speech recognition. In: Proc Wrkshp ASRU, IEEE (2011)
66. Senior, A., Lopez-Moreno, I.: Improving dnn speaker independence with i-vector inputs. In: Proc ICASSP, IEEE (2014) 225–229
67. Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., Khudanpur, S.: Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms. In: Proc Wrkshp ASRU, IEEE (2015) 539–546