

Towards Human Parity in Conversational Speech Recognition

Wayne Xiong, *Member, IEEE*, Jasha Droppo, *Senior Member, IEEE*, Xuedong Huang, *Fellow, IEEE*, Frank Seide, *Member, IEEE*, Mike Seltzer, *Senior Member, IEEE*, Andreas Stolcke, *Fellow, IEEE*, Dong Yu, *Senior Member, IEEE*, Geoffrey Zweig, *Fellow, IEEE*

Abstract—Conversational speech recognition has served as a flagship speech recognition task since the release of the Switchboard corpus in the 1990s. In this paper, we measure a human error rate on the widely used NIST 2000 test set for commercial bulk transcription. The error rate of professional transcribers is 5.9% for the Switchboard portion of the data, in which newly acquainted pairs of people discuss an assigned topic, and 11.3% for the CallHome portion where friends and family members have open-ended conversations. In both cases, our automated system edges past the human benchmark, achieving error rates of 5.8% and 11.0%, respectively. The key to our system’s performance is the use of various convolutional and long short-term memory acoustic model architectures, combined with a novel spatial smoothing method and lattice-free discriminative acoustic training, multiple recurrent neural network language modeling approaches, and a systematic use of system combination. Comparing frequent errors in our human and machine transcripts, we find them to be remarkably similar, and highly correlated as a function of the speaker. Human subjects find it very difficult to tell which errorful transcriptions come from humans. Overall, this suggests that, given sufficient matched training data, conversational speech transcription engines are approximating human parity in both quantitative and qualitative terms.

Index Terms—Conversational speech recognition, convolutional neural networks, recurrent neural networks, VGG, ResNet, LACE, BLSTM, spatial smoothing.

I. INTRODUCTION

Recent years have seen human performance levels reached or surpassed in tasks ranging from the games of chess and Go [1], [2] to simple speech recognition tasks like carefully read newspaper speech [3] and rigidly constrained small-vocabulary tasks in noise [4], [5]. In the area of speech recognition, much of the pioneering early work was driven by a series of carefully designed tasks with DARPA-funded data sets publicly released by the Linguistic Data Consortium (LDC) and the National Institute of Standards and Technology (NIST) [6]. At first, the recognition tasks were simple ones like Resource Management [7] with a small vocabulary and carefully controlled grammar; then read speech recognition in the Wall Street Journal task [8]; then Broadcast News [9]; each progressively more difficult for automatic systems. One of the last big initiatives in this area was in conversational telephone speech (CTS), which

is especially difficult due to the spontaneous (neither read nor planned) nature of the speech, its informality, and the self-corrections, hesitations and other disfluencies that are pervasive. The Switchboard [10] and later Fisher [11] data collections of the 1990s and early 2000s provide what is to date the largest and best studied of the conversational corpora. The history of work in this area includes key contributions by institutions such as IBM [12], BBN [13], SRI [14], AT&T [15], LIMSI [16], Cambridge University [17], Microsoft [18] and numerous others.

In the past, human performance on this task has been widely cited as being 4% [19]. However, the error rate estimate in [19] is attributed to a “personal communication,” and the actual source of this number is ephemeral. To better understand human performance, we have used professional transcribers to transcribe the actual test sets that we are working with, specifically the CallHome and Switchboard portions of the NIST eval 2000 test set. We find that the human error rates on these two parts are different almost by a factor of two, so a single number is inappropriate to cite. The error rate on Switchboard is about 5.9%, and for CallHome 11.3%. We improve on our recently reported conversational speech recognition system [20] by about 0.4%, and now exceed human performance by a small margin.

Our progress is a result of the careful engineering and optimization of convolutional and recurrent neural networks. While the basic structures have been well known for a long period [21], [22], [23], [24], [25], [26], [27], they have only recently started to dominate the field as the best models for speech recognition. This is the case for both acoustic modeling [28], [29], [30], [31], [32], [33] and language modeling [34], [35], [36], [37]. In comparison to the standard feed-forward multi-layer perceptrons (MLPs) or deep neural nets (DNNs) that first demonstrated breakthrough performance on conversational speech recognition [18], these acoustic models have the ability to model a large amount of acoustic context in the time dimension, and in the case of convolutional models, along the frequency dimension as well. In language modeling, recurrent models appear to improve over classical N-gram models through the use of an unbounded word history, as well as the generalization ability of continuous word representations [38], [39]. While ensemble classifier methods have long been popular in speech recognition systems [40], [41], [42] ensemble learning has now also become a feature in neural models [43], [44], [35], to improve robustness by reducing bias and variance.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer and A. Stolcke are with Microsoft AI and Research, 1 Microsoft Way, Redmond, WA 98052, U.S.A. D. Yu was with Microsoft Research, and is now with Tencent AI Lab. G. Zweig was with Microsoft Research, and is now with J.P. Morgan. Correspondence should be directed to andreas.stolcke@microsoft.com.

This paper is an expanded version of [20], with the following additional material:

- 1) A comprehensive discussion of human performance assessment for conversational speech transcription
- 2) The description of a novel spatial regularization method which significantly boosts our bidirectional long short-term memory (BLSTM) acoustic model performance
- 3) The use of long short-term memory (LSTM) rather than recurrent neural nets (RNNs) as language models, and the use of a letter-trigram input representation
- 4) A two-level system combination, based on a subsystem of BLSTM-system variants that, by itself, surpasses the best previously reported results
- 5) Expanded coverage of the Microsoft Cognitive Toolkit (CNTK) and hyperparameters used to build our models
- 6) An analysis of human versus machine error patterns and distribution (an expanded form of which appears in [45])

We note that this paper describes a snapshot of a system under development; an improved version is reported in [46].

The remainder of this paper describes our system in detail. Section II details our measurement of human performance and discusses related efforts. Section III describes the convolutional neural net (CNN) and long-short-term memory (LSTM) models. Section IV presents our lattice-free MMI training process. Language model rescoring is a significant part of our system, and described in Section V. We describe the CNTK toolkit that forms the basis of our neural network models, along with hyperparameters used, in Section VI. Experimental results are presented in Section VII, followed by an error analysis in Section VIII. We finish with a review of relevant prior and ongoing work in Section IX, followed by concluding remarks.

II. HUMAN PERFORMANCE

To measure human performance, we leveraged an existing pipeline in which Microsoft data is transcribed on a weekly basis. This pipeline uses a large commercial vendor to perform two-pass transcription. In the first pass, a transcriber works from scratch to transcribe the data. In the second pass, a second listener monitors the data to do error correction. Dozens of hours of test data are processed in each batch.

One week, we added the NIST 2000 CTS evaluation data to the work-list, without further comment. The intention was to measure the error rate of professional transcribers going about their normal everyday business. Aside from the standard two-pass checking in place, we did not do a complex multi-party transcription and adjudication process. The transcribers were given the same audio segments as were provided to the speech recognition system, which results in short sentences or sentence fragments from a single channel. This makes the task easier since the speakers are more clearly separated, and more difficult since the two sides of the conversation are not interleaved. Thus, it is the same condition as reported for our automated systems. The resulting numbers are 5.9% for the Switchboard portion, and 11.3% for the CallHome portion of the NIST 2000 test set, using the NIST scoring protocol. These numbers should be taken as an indication of the “error

rate” of a trained professional working in industry-standard speech transcript production. (We have submitted the human transcripts thus produced to the Linguistic Data Consortium for publication, so as to facilitate research by other groups.)

Note that a human transcriber error rate thus measured is really the “inter-annotator disagreement” between two human transcriptions, except that we identify one transcription as the “gold standard”, due the historical priority and greater level of effort involved in creating the NIST references. The disagreement between transcribers has been studied previously; researchers at the LDC [47] have assessed inter-transcriber error rates for data taken from the 2003 Rich Transcription evaluation set (which contains Switchboard and Fisher, but no CallHome data). Error rates of 4.1 to 4.5% are reported for extremely careful multiple transcriptions, and 9.6% for “quick transcriptions.” While this is a different test set, the numbers are in line with our findings.¹

Perhaps the most important point is the extreme variability between the two test subsets. The more informal CallHome data has almost double the human error rate of the Switchboard data. Interestingly, the same informality, multiple speakers per channel, and recording conditions that make CallHome hard for computers make it difficult for people as well. Notably, the performance of our artificial system aligns almost exactly with the performance of people on both sets.

We want to stress that the human results found by us are but one point on a continuum of possible levels of effort for human transcription, conveniently chosen to be indicative of an existing practice in the industry. Investing additional transcribers, effort, training, and possibly reconciliation between transcribers would all be expected to improve the final accuracy. The NIST study cited earlier represents a more careful transcription effort, albeit one that would probably be cost-prohibitive on a larger scale.

Another point of comparison is a recent human transcription experiment on the same data set that used a more involved transcription methodology [48]. In that work, the transcription vendor was cognizant of the experiment and explicitly familiarized with properties of the data and reference transcription conventions. Other important differences were that multiple independent first-pass transcribers were used, each of whom was making multiple listening passes over the data, and the overall best transcriber was chosen. Adding to that a quality assurance pass by a separate transcriber, between 12 and 18 listening passes were carried out. A final, and possibly crucial difference was that the transcribers in [48] could listen to the entire conversation side and thereby learn the characteristics of the speakers, their vocabulary, and conversation topic. This was not allowed in our pipeline since the vendor was set up to transcribe utterances independently. The human result in [48] were error rates of 5.1% and 6.8%, respectively, on Switchboard and CallHome. The improved result for CallHome transcription in particular could be due to the added context available to the transcribers.

¹We note that the bulk of the Fisher training data, and the bulk of the data overall, was transcribed with the “quick transcription” guidelines. Thus, the current state of the art is actually exceeding the human error level in the training data for the underlying models.

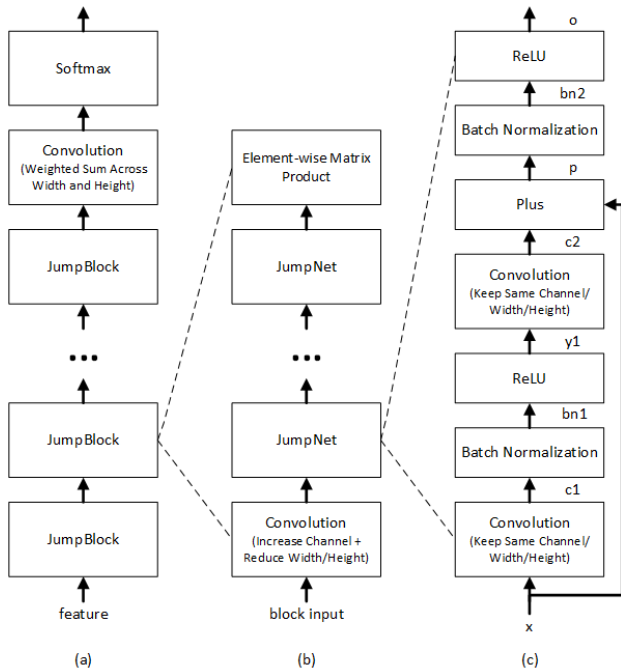


Fig. 1. LACE network architecture

III. CONVOLUTIONAL AND LSTM NEURAL NETWORKS

A. CNNs

We use three CNN variants. The first is the VGG architecture of [49]. Compared to the networks used previously in image recognition, this network uses small (3x3) filters, is deeper, and applies up to five convolutional layers before pooling.

The second network is modeled on the ResNet architecture [50], which adds highway connections [51], i.e. a linear transform of each layer’s input to the layer’s output [51], [52]. The only difference is that we apply batch normalization [53] before computing rectified linear unit (ReLU) activations.

The last CNN variant is the LACE (layer-wise context expansion with attention) model [54], depicted in Figure 1. LACE is a variant of time-delay neural network (TDNN) [23] in which each higher layer is a weighted sum of nonlinear transformations of a window of lower layer frames. In other words, each higher layer exploits broader context than the lower layers. Lower layers focus on extracting simple local patterns while higher layers extract complex patterns that cover broader contexts. Since not all frames in a window carry the same importance, a learned attention mask is applied, shown as the “element-wise matrix product” in Figure 1. The LACE model thus differs from the earlier TDNN models [23], [55] in this attention masking, as well as the ResNet-like linear pass-through. As shown in the diagram, the model is composed of four blocks, each with the same architecture. Each block starts with a convolution layer with stride two, which sub-samples the input and increases the number of channels. This layer is followed by four ReLU convolution layers with jump-links similar to those used in ResNet.

Table I compares the layer structure and parameters of the three CNN architectures. All three architectures employ batch

TABLE I
COMPARISON OF CNN ARCHITECTURES

	VGG Net	Residual-Net	LACE
Parameters	85M	38M	65M
Layers	14	49	22
Inputs	40x41	40x41	40x61
Convolution 1	[conv 3x3, 96] x3 maxpool (2x2)	[conv 1x1, 64 conv 3x3, 64 conv 1x1, 256] x3	jump block [conv 3x3, 128] x5
Convolution 2	[conv 3x3, 192] x4 maxpool (2x2)	[conv 1x1, 128 conv 3x3, 128 conv 1x1, 512] x4	jump block [conv 3x3, 256] x5
Convolution 3	[conv 3x3, 384] x4 maxpool (2x2)	[conv 1x1, 256 conv 3x3, 256 conv 1x1, 1024] x6	jump block [conv 3x3, 512] x5
Convolution 4		[conv 1x1, 512 conv 3x3, 512 conv 1x1, 2048] x3	jump block [conv 3x3, 1024] x5
Outputs	[FC 4096] x2 Softmax (9000)	average pool Softmax (9000)	[conv 3x4, 1] x1 Softmax (9000)

normalization [53] between layers.

We also trained a fused model by combining a ResNet model and a VGG model at the senone posterior level. (A senone represents a tied context-dependent HMM state.) Both base models are independently trained, and then the score fusion weight is optimized on development data. The fused system is our best single system.

B. LSTMs

While our best performing models are convolutional, the use of long short-term memory networks (LSTMs) is a close second. We use a bidirectional architecture [56] without frame-skipping [29]. The core model structure is the LSTM defined in [28]. We found that using networks with more than six layers did not improve the word error rate on the development set, and chose 512 hidden units, per direction, per layer, as that provided a reasonable trade-off between training time and final model accuracy.

C. Spatial Smoothing

Inspired by the human auditory cortex [57], where neighboring neurons tend to simultaneously activate, we employ a spatial smoothing technique to improve the accuracy of our LSTM models. We make no claim about the neural plausibility of the resulting model, and view the imposition of a topological smoothing constraint simply as a device for regularizing network parameter estimation. Stimulated networks [58] employ a similar approach.

The smoothing is implemented as a regularization term on the activations between layers of the acoustic model. First, each vector of activations is re-interpreted as a 2-dimensional image. For example, if there are 512 neurons, they are interpreted as the pixels of a 16 by 32 image. Second, this image is high-pass filtered. The filter is implemented as a circular convolution with a 3 by 3 kernel. The center tap of the kernel has a value of 1, and the remaining eight having a value of $-1/8$. Third, the energy (L2 norm) of this high-pass filtered image is computed and added to the training objective function. We have found empirically that a

TABLE II
ACCURACY IMPROVEMENTS FROM SPATIAL SMOOTHING ON THE NIST 2000 CTS TEST SET. THE MODEL IS A SIX LAYER BLSTM, USING I-VECTORS AND 40 DIMENSIONAL FILTERBANK FEATURES, AND A DIMENSION OF 512 IN EACH DIRECTION OF EACH LAYER. SENONES REFERS TO THE NUMBER OF TIED CONTEXT-DEPENDENT HMM STATES.

Model	WER (%)			
	9000 senones		27000 senones	
	CH	SWB	CH	SWB
Baseline	21.4	9.9	20.5	10.6
Spatial smoothing	19.2	9.3	19.5	9.2

suitable scale for this energy is 0.1 with respect to the existing cross entropy objective function. We experimented with 3-dimensional smoothing topologies as well, but found them to give worse results than the 2-dimensional arrangement.

The overall effect of this process is to make the training algorithm prefer models that have correlated neurons, and to improve the word error rate of the acoustic model. Table II shows the benefit in error rate for some of our early systems. We observed error reductions of between 5 and 10% relative from spatial smoothing. Note that without spatial smoothing, increasing the number of senones has a detrimental effect on SWB accuracy. However, spatial smoothing remedies this anomaly, suggesting it was due to overfitting, or insufficient regularization.

D. Speaker Adaptive Modeling

Speaker adaptive modeling in our system is based on conditioning the network on an i-vector [59] characterization of each speaker [60], [61]. A 100-dimensional i-vector is generated for each conversation side (channel A or B of the audio file, i.e., all the speech coming from the same speaker). For the LSTM system, the conversation-side i-vector v_s is appended to each frame of input. For convolutional networks, this approach is inappropriate because we do not expect to see spatially contiguous patterns in the input. Instead, for the CNNs, we add a learnable weight matrix W^l to each layer, and add $W^l v_s$ to the activation of the layer before the nonlinearity. Thus, in the CNN, the i-vector essentially serves as an speaker-dependent bias to each layer. Note that the i-vectors are estimated using MFCC features; by using them subsequently in systems based on log-filterbank features, we may benefit from a form of feature combination.

Performance improvements from i-vectors are shown in Table III. The full experimental setup is described in Section VII. We note that the BLSTM model on CH data (but not the other models or for SWB data) shows a slight degradation under i-vector adaptation. We conjecture that is due to the fact that some CH conversations have multiple speakers on one conversation side (see Section VIII). Since the BLSTM uses information over longer time spans this could create a mismatch around speaker change points.

IV. LATTICE-FREE SEQUENCE TRAINING

After standard cross-entropy training, we optimize the model parameters using the maximum mutual information

(MMI) objective function. Denoting a word sequence by w and its corresponding acoustic realization by a , the training criterion is

$$\sum_{w,a \in \text{data}} \log \frac{P(w)P(a|w)}{\sum_{w'} P(w')P(a|w')}$$

As noted in [62], [63], the necessary gradient for use in backpropagation is a simple function of the posterior probability of a particular acoustic model state at a given time, as computed by summing over all possible word sequences in an unconstrained manner. As first done in [12], and more recently in [64], this can be accomplished with a straightforward alpha-beta computation over the finite state acceptor representing the decoding search space. In [12], the search space is taken to be an acceptor representing the composition $HCLG$ for a unigram language model L on words. In [64], a language model on phonemes is used instead.

In our implementation, we use a mixed-history acoustic unit language model. In this model, the probability of transitioning into a new context-dependent phonetic state (senone) is conditioned on both the senone and phone history. We found this model to perform better than either purely word-based or phone-based models. Based on a set of initial experiments, we developed the following procedure:

- 1) Perform a forced alignment of the training data to select lexical variants and determine frame-aligned senone sequences.
- 2) Compress consecutive framewise occurrences of a single senone into a single occurrence.
- 3) Estimate an unsmoothed, variable-length N-gram language model from this data, where the history state consists of the previous phone and previous senones within the current phone.

To illustrate this, consider the sample senone sequence $\{s_s2.1288, s_s3.1061, s_s4.1096\}$, $\{eh_s2.527, eh_s3.128, eh_s4.66\}$, $\{t_s2.729, t_s3.572, t_s4.748\}$. When predicting the state following $eh_s4.66$ the history consists of $(s, eh_s2.527, eh_s3.128, eh_s4.66)$, and following $t_s2.729$, the history is $(eh, t_s2.729)$.

We construct the denominator graph from this language model, and HMM transition probabilities as determined by transition-counting in the senone sequences found in the training data. Our approach not only largely reduces the complexity of building up the language model but also provides very reliable training performance.

We have found it convenient to do the full computation, without pruning, in a series of matrix-vector operations on the GPU. The underlying acceptor is represented with a sparse matrix, and we maintain a dense likelihood vector for each time frame. The alpha and beta recursions are implemented with CUSPARSE level-2 routines: sparse-matrix, dense vector multiplies. Run time is about 100 times faster than real time. As in [64], we use cross-entropy regularization. In all the lattice-free MMI (LFMMI) experiments mentioned below we use a trigram language model over senone and phone labels, as explained earlier. Most of the gain is usually obtained after processing 24 to 48 hours of data.

TABLE III
PERFORMANCE IMPROVEMENTS FROM I-VECTOR AND LFMMI TRAINING ON THE NIST 2000 CTS TEST SET

Configuration	WER (%)							
	ReLU-DNN		ResNet-CNN		BLSTM		LACE	
	CH	SWB	CH	SWB	CH	SWB	CH	SWB
Baseline	21.9	13.4	17.5	11.1	17.3	10.3	16.9	10.4
i-vector	20.1	11.5	16.6	10.0	17.6	9.9	16.4	9.3
i-vector+LFMMI	17.9	10.2	15.2	8.6	16.3	8.9	15.2	8.5

V. LM RESCORING AND SYSTEM COMBINATION

An initial decoding is done with a WFST decoder, using the architecture described in [65]. We use an N-gram language model trained and pruned with the SRILM toolkit [66]. The first-pass LM has approximately 15.9 million bigrams, trigrams, and 4grams, and a vocabulary of 30,500 words. It gives a perplexity of 69 on the 1997 CTS evaluation transcripts. The initial decoding produces a lattice with the pronunciation variants marked, from which 500-best lists are generated for rescoring purposes. Subsequent N-best rescoring uses an unpruned LM comprising 145 million N-grams. All N-gram LMs were estimated by a maximum entropy criterion as described in [67].

The N-best hypotheses are then rescored using a combination of the large N-gram LM and several neural net LMs. We have experimented with both RNN LMs and LSTM LMs, and describe the details in the following two sections. Note we delay application of recurrent neural net LMs, which use an unlimited history, to the N-best rescoring phase as a matter of convenience, since complete sentence hypotheses are available at that point.

We assessed the lower bound of performance for our lattice/N-best rescoring paradigm. The 500-best lists from the lattices generated with the ResNet CNN system had an oracle (lowest achievable) WER of 2.7% on the Switchboard portion of the NIST 2000 evaluation set, and an oracle WER of 4.9% on the CallHome portion. The oracle error of the combined system is even lower (though harder to quantify) since (1) N-best output from all systems are combined and (2) confusion network construction generates new possible hypotheses not contained in the original N-best lists. With oracle error rates less than half the currently achieved actual error rates (as reported below), we conclude that search errors are not a major limiting factor in the rescoring framework.

A. RNN-LM setup

Our RNN-LMs are trained and evaluated using the CUED-RNNLM toolkit [68]. Our RNN-LM configuration has several distinctive features, as described below.

- 1) We trained both standard, forward-predicting RNN-LMs and backward RNN-LMs that predict words in reverse temporal order. The log probabilities from both models are added.
- 2) As is customary, the RNN-LM probability estimates are interpolated at the word-level with corresponding N-gram LM probabilities (separately for the forward and backward models). In addition, we trained a second

RNN-LM for each direction, obtained by starting with different random initial weights. The two RNN-LMs and the N-gram LM for each direction are interpolated with weights of (0.375, 0.375, 0.25).

- 3) In order to make use of LM training data that is not fully matched to the target conversational speech domain, we start RNN-LM training with the union of in-domain (here, CTS) and out-of-domain (e.g., Web) data. Upon convergence, the network undergoes a second training phase using the in-domain data only. Both training phases use in-domain validation data to regulate the learning rate schedule and termination. Because the size of the out-of-domain data is a multiple of the in-domain data, a standard training on a simple union of the data would not yield a well-matched model, and have poor perplexity in the target domain.
- 4) We found best results with an RNN-LM configuration that had a second, non-recurrent hidden layer. This produced lower perplexity and word error than the standard, single-hidden-layer RNN-LM architecture [34].² The overall network architecture thus had two hidden layers with 1000 units each, using ReLU nonlinearities. Training used noise-contrastive estimation (NCE) [69].
- 5) The RNN-LM output vocabulary consists of all words occurring more than once in the in-domain training set. While the RNN-LM estimates a probability for unknown words, we take a different approach in rescoring: The number of out-of-set words is recorded for each hypothesis and a penalty is estimated for them when optimizing the relative weights for all model scores (acoustic, LM, pronunciation), using the SRILM *nbest-optimize* tool.

B. LSTM-LM setup

After obtaining good results with RNN-LMs we also explored the LSTM recurrent network architecture for language modeling, inspired by recent work showing gains over RNN-LMs for conversational speech recognition [37]. In addition to applying the lessons learned from our RNN-LM experiments, we explored additional alternatives, as described below.

- 1) The LSTM LMs take one two types of input vectors: word based one-hot vector input and letter trigram vector [70] input. Including both forward and backward models, we trained four different LSTM LMs in total.
- 2) For the word based input, we adopted the approach from [71] to tie the input embedding and output embedding together.

²However, adding more hidden layers produced no further gains.

TABLE IV
LSTM PERPLEXITIES (PPL) AS A FUNCTION OF HIDDEN LAYERS, TRAINED ON IN-DOMAIN DATA ONLY, COMPUTED ON 1997 CTS EVAL TRANSCRIPTS.

Language model	PPL
letter trigram input with one layer (baseline)	63.2
+ two hidden layers	61.8
+ three hidden layers	59.1
+ four hidden layers	59.6
+ five hidden layers	60.2
+ six hidden layers	63.7

TABLE V
PERPLEXITIES (PPL) OF THE FOUR LSTM LMS USED IN THE FINAL COMBINATION. PPL IS COMPUTED ON 1997 CTS EVAL TRANSCRIPTS. ALL THE LSTM LMS ARE WITH THREE HIDDEN LAYERS.

Language model	PPL
RNN: 2 layers + word input (baseline)	59.8
LSTM: word input in forward direction	54.4
LSTM: word input in backward direction	53.4
LSTM: letter trigram input in forward direction	52.1
LSTM: letter trigram input in backward direction	52.0

- 3) Here we also used a two-phase training schedule to train the LSTM LMs. First we train the model on the combination of in-domain and out-domain data for four data passes without any learning rate adjustment. We then start from the resulting model and train on in-domain data until convergence.
- 4) Overall the letter-trigram-based models perform a little better than the word-based language model. The letter trigram representation had dimension 7190 (the number of distinct trigrams in the vocabulary), compared to a learned embedding of size 1000 for the word-based model. (Note that no extra embedding layer was included for letter trigram inputs.)
- 5) Convergence was improved through a variation of self-stabilization [72], in which each output vector x of nonlinearities are scaled by $\frac{1}{4} \ln(1 + e^{4\beta})$, where a β is a scalar that is learned for each output. This has a similar effect as the scale of the well-known batch normalization technique [53], but can be used in recurrent loops.
- 6) Table IV shows the impact of number of layers on the final perplexities. Based on this, we proceeded with three hidden layers, with 1000 hidden units each. The perplexities of each LSTM-LM we used in the final combination (before interpolating with the N-gram model) can be found in Table V.

For the final system, we interpolated two LSTM-LMs with an N-gram LM for the forward-direction LM, and similarly for the backward-direction LM. All LSTMs use three hidden layers and are trained on in-domain and web data. Unlike for the RNN-LMs, the two models being interpolated differ not just in their random initialization, but also in the input encoding (one uses a triletter encoding, the other a one-hot word encoding). The forward and backward LM log probability scores are combined additively.

TABLE VI
PERFORMANCE OF VARIOUS VERSIONS OF NEURAL-NET-BASED LM RESCORING. PERPLEXITIES (PPL) ARE COMPUTED ON 1997 CTS EVAL TRANSCRIPTS; WORD ERROR RATES (WER) ON THE NIST 2000 SWITCHBOARD TEST SET.

Language model	PPL	WER
4-gram LM (baseline)	69.4	8.6
+ RNNLM, CTS data only	62.6	7.6
+ Web data training	60.9	7.4
+ 2nd hidden layer	59.0	7.4
+ 2-RNNLM interpolation	57.2	7.3
+ backward RNNLMs	-	6.9
+ LSTM-LM, CTS + Web data	51.4	6.9
+ 2-LSTM-LM interpolation	50.5	6.8
+ backward LSTM-LM	-	6.6

C. Training data

The 4-gram language model for decoding was trained on the available CTS transcripts from the DARPA EARS program: Switchboard (3M words), BBN Switchboard-2 transcripts (850k), Fisher (21M), English CallHome (200k), and the University of Washington conversational Web corpus (191M). A separate N-gram model was trained from each source and interpolated with weights optimized on RT-03 transcripts. For the unpruned large rescoring 4-gram, an additional LM component was added, trained on 133M word of LDC Broadcast News texts. The N-gram LM configuration is modeled after that described in [61], except that maxent smoothing was used.

The RNN and LSTM LMs were trained on Switchboard and Fisher transcripts as in-domain data (20M words for gradient computation, 3M for validation). To this we added 62M words of UW Web data as out-of-domain data, for use in the two-phase training procedure described above.

D. RNN-LM and LSTM-LM performance

Table VI gives perplexity and word error performance for various recurrent neural net LM setups, from simple to more complex. The acoustic model used was the ResNet CNN. Note that, unlike the results in Tables IV and V, the neural net LMs in Table VI are interpolated with the N-gram LM. As found by others [73], LSTMs outperform corresponding RNN-LMs.

As can be seen, each of the measures described earlier adds incremental gains, which, however, add up to a substantial improvement overall. The total gain relative to a purely N-gram based system is a 20% relative error reduction with RNN-LMs, and 23% with LSTM-LMs. As shown later (see Table VIII) the gains with different acoustic models are similar.

E. System Combination

The LM rescoring is carried out separately for each acoustic model. The rescored N-best lists from each subsystem are then aligned into a single confusion network [42] using the SRILM *nbest-rover* tool. This method is a form of confusion network-based system combination that does not require building separate confusion networks for the individual subsystems. However, the number of potential candidate systems is too large to allow an all-out combination, both for practical

reasons and due to overfitting issues. Instead, we perform a greedy search, starting with the single best system, and successively adding additional systems, to find a small set of systems that are maximally complementary. The RT-02 Switchboard set was used for this search procedure.

We experimented with two search algorithms to find good subsets of systems. We always start with the system giving the best individual accuracy on the development set. In one approach, a greedy forward search then adds systems incrementally to the combination, giving each equal weight. If no improvement is found with any of the unused systems, we try adding each with successively lower relative weights of 0.5, 0.2, and 0.1, and stop if none of these give an improvement.

A second variant of the search procedure that can give lower error (as measured on the devset) estimates the best system weights for each incremental combination candidate. The weight estimation is done using an expectation-maximization algorithm based on aligning the reference words to the confusion networks, and maximizing the weighted probability of the correct word at each alignment position. To avoid overfitting, the weights for an N -way combination are smoothed hierarchically, i.e., interpolated with the weights from the $(N - 1)$ -way system that preceded it. This tends to give robust weights that are biased toward the early (i.e., better) subsystems.

The final system incorporated a variety of BLSTM models with roughly similar performance, but differing in various metaparameters (number of senones, use of spatial smoothing, and choice of pronunciation dictionaries).³ To further limit the number of free parameters to be estimated in system combination, we performed system selection in two stages. First, we selected the four best BLSTM systems. We then combined these with equal weights and treated them as a single subsystem in searching for a larger combination including other acoustic models. This yielded our best overall combined system, as reported in Section VII-C.

VI. MICROSOFT COGNITIVE TOOLKIT (CNTK)

All neural networks in the final system were trained with the Microsoft Cognitive Toolkit, or CNTK [74], [75] on a Linux-based multi-GPU server farm. CNTK allows for flexible model definition, while at the same time scaling very efficiently across multiple GPUs and multiple servers. The resulting fast experimental turnaround using the full 2000-hour corpus was critical for our work.

A. Flexible, Terse Model Definition

In CNTK, a neural network (and the training criteria) are specified by its formula, using a custom functional language (BrainScript), or Python. A graph-based execution engine, which provides automatic differentiation, then trains the model's parameters through stochastic gradient descent (SGD). Leveraging a stock library of common layer types, networks can be specified very tersely. Samples can be found in [74].

³We used two different dictionaries, one based on a standard phone set and another with dedicated vowel and nasal phones used only in the pronunciations of filled pauses (“uh”, “um”) and backchannel acknowledgments (“uh-huh”, “mhm”), similar to [42].

B. Multi-Server Training using 1-bit SGD

Training the acoustic models in this paper on a single GPU would take many weeks or even months. CNTK made training times feasible by parallelizing the SGD training with our 1-bit SGD parallelization technique [76]. This data-parallel method distributes minibatches over multiple worker nodes, and then aggregates the sub-gradients. While the necessary communication time would otherwise be prohibitive, the 1-bit SGD method eliminates the bottleneck by two techniques: 1-bit quantization of gradients and automatic minibatch-size scaling.

In [76], we showed that gradient values can be quantized to just a single bit, if one carries over the quantization error from one minibatch to the next. Each time a sub-gradient is quantized, the quantization error is computed and remembered, and then added to the next minibatch's sub-gradient. This reduces the required bandwidth 32-fold with minimal loss in accuracy. Secondly, automatic minibatch-size scaling progressively decreases the frequency of model updates. At regular intervals (e.g. every 72h of training data), the trainer tries larger minibatch sizes on a small subset of data and picks the largest that maintains training loss.

These two techniques allow for excellent multi-GPU/server scalability, and reduced the acoustic-model training times on 2000h from months to between 1 and 3 weeks, making this work feasible.

C. Training hyperparameters

As mentioned, all our models are trained in parallelized fashion with CNTK and 1-bit SGD enabled. We use the “FsAdaGrad” learning algorithm, which is an implementation of Adam [77] in CNTK. A typical learning rate is 3×10^{-6} , and learning rates are automatically adjusted with a decrease factor of 0.7. Momentum is set at a constant value of 2500 throughout model training. For individual acoustic models, we find that training converges after 1.5 to 2 passes over the 2000-hour training set.

We do not use dropout or gradient noise in our model training; an alternative regularization method that we apply only to BLSTM model training is the spatial smoothing described in Section III-C.

D. Computational performance

Table VII compares the runtimes, as multiples of speech duration, of various processing steps associated with the different acoustic model architectures (figures for DNNs are given only as a reference point, since they are not used in our system). Acoustic model (AM) training comprises the forward and backward dynamic programming passes, as well as parameter updates. AM evaluation refers to the forward computation only. Decoding includes AM evaluation along with hypothesis search (only the former makes use of the GPU). Runtimes were measured on a 12-core Intel Xeon E5-2620v3 CPU clocked at 2.4GHz, with an Nvidia Titan X GPU. We observe that the GPU gives a 10 to 100-fold speedup for AM evaluation over the CPU implementation. AM evaluation is thus reduced to a small fraction of overall decoding time, making near-realtime operation possible.

TABLE VII
 RUNTIMES AS FACTOR OF SPEECH DURATION FOR VARIOUS ASPECTS OF ACOUSTIC MODELING AND DECODING, FOR DIFFERENT TYPES OF ACOUSTIC MODEL

Processing step	Hardware	DNN	ResNet-CNN	BLSTM	LACE
AM training	GPU	0.012	0.60	0.022	0.23
AM evaluation	GPU	0.0064	0.15	0.0081	0.081
AM evaluation	CPU	0.052	11.7	n/a	8.47
Decoding	GPU	1.04	1.19	1.40	1.38

VII. EXPERIMENTS AND RESULTS

A. Speech corpora

We train with the commonly used English CTS (Switchboard and Fisher) corpora. Evaluation is carried out on the NIST 2000 CTS test set, which comprises both Switchboard (SWB) and CallHome (CH) subsets. The waveforms were segmented according to the NIST partitioned evaluation map (PEM) file, but the cut points specified could leave too little nonspeech audio around the edges, interfering with feature extraction windows. This seemed to cause problems with the CallHome data, and we therefore padded the PEM segments with 150ms of dithered silence.⁴ The Switchboard-1 portion of the NIST 2002 CTS test set was used for tuning and development. The acoustic training data is comprised by LDC corpora 97S62, 2004S13, 2005S13, 2004S11 and 2004S09; see [12] for a full description.

B. Acoustic Model Details

Forty-dimensional log-filterbank features were extracted every 10 milliseconds, using a 25-millisecond analysis window. The CNN models used window sizes as indicated in Table I, and the LSTMs processed one frame of input at a time. The bulk of our models use three state left-to-right triphone models with 9000 tied states. Additionally, we have trained several models with 27k tied states. The phonetic inventory includes special models for noise, vocalized-noise, laughter and silence. We use a 30k-vocabulary derived from the most common words in the Switchboard and Fisher corpora. The decoder uses a statically compiled unigram graph, and dynamically applies the language model score. The unigram graph has about 300k states and 500k arcs.

Table III shows the result of i-vector adaptation and LFMMI training on several of our early systems. We achieve a 5–8% relative improvement from i-vectors, including on CNN systems. The last row of Table III shows the effect of LFMMI training on the different models. We see a consistent 7–10% further relative reduction in error rate for all models. Considering the great increase in procedural simplicity of LFMMI over the previous practice of writing lattices and post-processing them, we consider LFMMI to be a significant advance in technology.

C. Overall Results and Discussion

The performance of all our component models is shown in Table VIII, along with the BLSTM combination, the full system combination results, and the measured human transcriber

error. (Recall that the four best BLSTM systems are combined with equal weights first, as described in Section V-E.)⁵

Table VIII lists all models that are selected for the combined systems, for one or more of the three rescoring LMs. The only exception is the VGG+ResNet system, which combines acoustic senone posteriors from the VGG and ResNet networks. While this yields a system that beats the individual acoustic models, only the individual VGG and ResNet models are used in the overall system combination.

We also observe that the four model variants chosen for the combined BLSTM subsystem differ incrementally by one hyperparameter (smoothing, number of senones, dictionary), and that the BLSTMs alone achieve an error that is within 3% relative of the full system combination. This validates the rationale that choosing different hyperparameters is an effective way to obtain complementary systems for combination purposes.

VIII. ERROR ANALYSIS

In this section, we compare the errors made by our artificial recognizer with those made by human transcribers. First we note that machine and human transcription WERs do not differ significantly according to the Wilcoxon and Matched Pairs Sentence Segment Word Error tests as applied by NIST, nor do they differ according to a Sign test comparing error counts at the utterance level.

A first high-level question regarding the relation between word errors by machine and human transcribers is whether difficulty in one predicts difficulty in the other. Figure 2 shows scatter plots of speaker-level error rates (machine vs. human), separated by corpus. Each corpus subset has 40 conversation sides.

Clearly the errors at that level are correlated, with $\rho = 0.65$ for SWB and $\rho = 0.73$ for CH. This suggests that properties of the speech, either as a function of the content, the speaker, or the channel (each speaker occurs in exactly one test conversation), cause errors for both machine and human transcription.

We observe that the CH data has two speakers with outlier machine error rates (37.5% and 64.7% WER, solid red dots in Figure 2). These correspond to secondary speakers in their respective conversation sides, each with only a fraction of the speech of the dominant speaker. Note that the ASR system processes each conversation assuming only a single speaker

⁵When comparing the last row in Table III with the “N-gram LM” results in Table VIII, note that the former results were obtained with the pruned N-gram LM used in the decoder and fixed score weights (during lattice generation), whereas the latter results are from rescoring with the unpruned N-gram LM (during N-best generation), using optimized score weighting. Accordingly, the rescoring results are generally somewhat better.

⁴Using the `sox` tool options `pad 0.15 0.15 dither -p 14`.

TABLE VIII

WORD ERROR RATES (%) ON THE NIST 2000 CTS TEST SET WITH DIFFERENT ACOUSTIC MODELS, AND HUMAN ERROR RATE FOR COMPARISON. UNLESS OTHERWISE NOTED, MODELS ARE TRAINED ON THE FULL 2000 HOURS OF DATA AND HAVE 9K SENONES. THE THIRD RESNET SYSTEM WAS TRAINED ON GMM-BASED ALIGNMENTS, SKIPPING THE REALIGNMENT WITH CNN MODELS.

Model	N-gram LM		RNN-LM		LSTM-LM	
	CH	SWB	CH	SWB	CH	SWB
ResNet, 300h training	19.2	10.0	17.7	8.2	17.0	7.7
ResNet	14.8	8.6	13.2	6.9	12.5	6.6
ResNet, GMM alignments	15.3	8.8	13.7	7.3	12.8	6.9
VGG	15.7	9.1	14.1	7.6	13.2	7.1
VGG + ResNet	14.5	8.4	13.0	6.9	12.2	6.4
LACE	15.0	8.4	13.5	7.2	13.0	6.7
BLSTM	16.5	9.0	15.2	7.5	14.4	7.0
BLSTM, spatial smoothing	15.4	8.6	13.7	7.4	13.0	7.0
BLSTM, spatial smoothing, 27k senones	15.3	8.3	13.8	7.0	13.2	6.8
BLSTM, spatial smoothing, 27k senones, alternate dictionary	14.9	8.3	13.7	7.0	13.0	6.7
BLSTM system combination	13.2	7.3	12.1	6.4	11.6	6.0
Full system combination	13.0	7.3	11.7	6.1	11.0	5.8
Human transcription					11.3	5.9

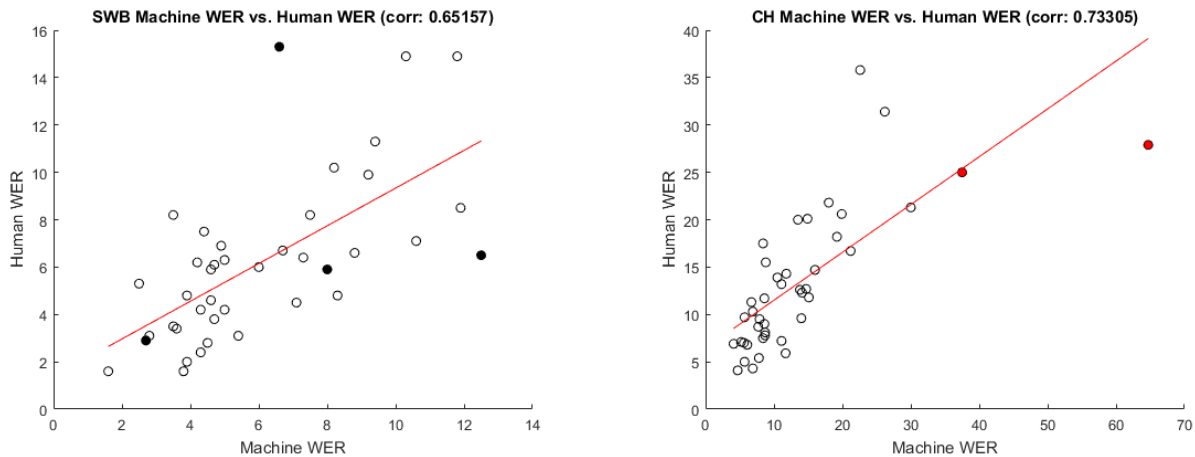


Fig. 2. Correlation between machine and human word error rates at speaker level. The solid black circles represent SWB speakers *not* seen in training. The solid red circles stand for secondary CH speakers that share a conversation side with a dominating primary speaker.

per side. If we remove these outliers, the machine-human error correlation on CH increases to $\rho = 0.80$. With secondary speakers excluded, we can also observe that the machine error rates cluster tighter than the human ones in both corpora (SWB: machine $6.1\% \pm 2.8$ vs. human $6.2\% \pm 3.4$; CH: machine $11.8\% \pm 6.1$ vs. human $12.7\% \pm 7.0$).

In [48] it was suggested that one of the reasons for the much higher error rate on CH compared to SWB was that 36 of the 40 SWB test speakers occur in the portion of the SWB corpus that is used in training. To assess this hypothesis we singled out the four speakers in the SWB portion that are not found in the training set; these are shown as solid black circles in Figure 2. At first, it seems that the speaker-averaged WER for the “seen” speakers (machine WER 5.9%) is indeed much lower than for the speakers not found in training (7.5%). However, we can attribute this to bad luck and small sample size, since the average machine WER of 7.5% for “unseen” speakers is well within one standard deviation of the “seen” speakers’ WER distribution ($5.9\% \pm 2.7$), and more tellingly, almost exactly the same relative difference in WERs between “seen” and “unseen” speakers is observed for human transcriptions (6.0%

versus 7.7%). Clearly the human transcribers did not have the benefit of training on the “seen” speakers, so the difference must be due to the intrinsic difficulty of the speakers, which affects both transcription systems.

Table IX shows the ten most common substitutions for both humans and the machine transcription system. Tables X and XI do the same for deletions and insertions. We see an overall similarity in human and machine error types, in that short function words and discourse particles dominate the rankings. These words are plausible as top error type for two reasons: they occur frequently, and are often realized in a phonetically reduced manner and are therefore confusable [78].

Focusing on the substitutions, we see that by far the most common error in the ASR system is the confusion of a *hesitation* in the reference for a *backchannel acknowledgment* (“%bckack” after NIST text normalization) in the hypothesis. Human transcribers make this mistake too, but much less frequently. The pragmatic significance is that backchannel words like “uh-huh” are an acknowledgment of the speaker, also signaling that the speaker should keep talking, while hesitations like “uh” and “um” are used to indicate that the

TABLE IX
MOST COMMON SUBSTITUTIONS FOR ASR SYSTEM AND HUMANS. THE NUMBER OF TIMES EACH ERROR OCCURS IS FOLLOWED BY THE WORD IN THE REFERENCE, AND WHAT APPEARS IN THE HYPOTHESIS INSTEAD.

CH		SWB	
ASR	Human	ASR	Human
45: (%hesitation) / %bcack	12: a / the	29: (%hesitation) / %bcack	12: (%hesitation) / hmm
12: was / is	10: (%hesitation) / a	9: (%hesitation) / oh	10: (%hesitation) / oh
9: (%hesitation) / a	10: was / is	9: was / is	9: was / is
8: (%hesitation) / oh	7: (%hesitation) / hmm	8: and / in	8: (%hesitation) / a
8: a / the	7: bentsy / bensì	6: (%hesitation) / i	5: in / and
7: and / in	7: is / was	6: in / and	4: (%hesitation) / %bcack
7: it / that	6: could / can	5: (%hesitation) / a	4: and / in
6: in / and	6: well / oh	5: (%hesitation) / yeah	4: is / was
5: a / to	5: (%hesitation) / %bcack	5: a / the	4: that / it
5: aw / oh	5: (%hesitation) / oh	5: jeez / jeeze	4: the / a

TABLE X
MOST COMMON DELETIONS FOR ASR SYSTEM AND HUMANS.

CH		SWB	
ASR	Human	ASR	Human
44: i	73: i	31: it	34: i
33: it	59: and	26: i	30: and
29: a	48: it	19: a	29: it
29: and	47: is	17: that	22: a
25: is	45: the	15: you	22: that
19: he	41: %bcack	13: and	22: you
18: are	37: a	12: have	17: the
17: oh	33: you	12: oh	17: to
17: that	31: oh	11: are	15: oh
17: the	30: that	11: is	15: yeah

TABLE XI
MOST COMMON INSERTIONS FOR ASR SYSTEM AND HUMANS.

CH		SWB	
ASR	Human	ASR	Human
15: a	10: i	19: i	12: i
15: is	9: and	9: and	11: and
11: i	8: a	7: of	9: you
11: the	8: that	6: do	8: is
11: you	8: the	6: is	6: they
9: it	7: have	5: but	5: do
7: oh	5: you	5: yeah	5: have
6: and	4: are	4: air	5: it
6: in	4: is	4: in	5: yeah
6: know	4: they	4: you	4: a

current speaker has more to say and wants to keep his or her turn.⁶ As turn-management devices, these two classes of words therefore have exactly opposite functions, and come with prosodic and contextual cues that a human listener would use for recognition. For the machine, these words could be more confusable due to their phonetic similarity, and absence of other cues. Inconsistent labeling in the training data, the bulk

⁶The NIST scoring protocol treats hesitation words as optional, and we therefore delete them from our recognizer output prior to scoring. This explains why we see many substitutions of backchannels for hesitations, but not vice-versa.

TABLE XII
OVERALL SUBSTITUTION, DELETION AND INSERTION RATES.

	CH		SWB	
	ASR	Human	ASR	Human
sub	6.5	4.1	3.3	2.6
del	3.3	6.5	1.8	2.7
ins	1.4	0.7	0.7	0.7
all	11.1	11.3	5.9	5.9

of which comes from “quick transcription” [47] of the Fisher corpus, could also contribute to the frequency of this error type. For perspective, however, there are over 21,000 words in each test set. Thus the hesitation/backchannel substitutions contribute only about 0.2% absolute to the overall error rate.

The most frequent substitution for humans on the Switchboard corpus was mistaking a hesitation in the reference for the word “hmm.” The scoring guidelines treat “hmm” as a word distinct from backchannels and hesitations, so this is not a scoring mistake. Examination of the contexts in which the error is made show that it is most often intended to acknowledge the other speaker, i.e. as a backchannel. For both people and our automated system, the insertion and deletion patterns are similar: short function words are by far the most frequent errors. In particular, the single most common error made by the transcribers was to omit the word “I.” While we believe further improvement in function and content words is possible, the significance of the remaining backchannel/hesitation confusions is unclear.

Table XII shows the overall error rates broken down by substitutions, insertions and deletions. We see that the human transcribers have a somewhat lower substitution rate, and a higher deletion rate. The relatively higher deletion rate might reflect a human bias to avoid outputting uncertain information, or the productivity demands on a professional transcriber. In all cases, the number of insertions is relatively small.

As a final investigation into the qualitative differences between human and machine errors we conducted an informal human subject experiment, described in more detail elsewhere [45]. We presented participants with the NIST reference transcripts, our own human transcripts, and the machine recog-

dition outputs, side by side. The data was presented one utterance at a time, randomly sampled from among utterances where humans and machine differed. Human and machine transcripts were not identified as such and arranged randomly, and the subject was asked to identify the human transcript. After each trial the subject learned whether they were correct or not, and moved on to the next trial. Overall, participants achieved a success rate of 53%, not significantly different from chance. While only an informal experiment, it demonstrated that systematic differences between human and machine output are quite hard to discern by people for this type of speech.

IX. RELATION TO PRIOR AND ONGOING WORK

Compared to earlier applications of CNNs to speech recognition [79], [80], our networks are much deeper, and use linear bypass connections across convolutional layers. They are similar in spirit to those studied more recently by [31], [30], [61], [32], [33]. We improve on these architectures with the LACE model [54], which iteratively expands the effective window size, layer-by-layer, and adds an attention mask to differentially weight distant context.

Our spatial regularization technique is similar in spirit to stimulated deep neural networks [58]. Whereas stimulated networks use a supervision signal to encourage locality of activations in the model, our technique is automatic.

Our use of lattice-free MMI is distinctive, and extends previous work [12], [64] by proposing the use of a mixed triphone/phoneme history in the language model.

On the language modeling side, we achieve a performance boost by combining multiple LSTM-LMs in both forward and backward directions, and by using a two-phase training regimen to get best results from out-of-domain data. For our best CNN system, LSTM-LM rescoring yields a relative word error reduction of 23%, and a 20% relative gain for the combined recognition system, considerably larger than previously reported for conversational speech recognition [37].

Conversational speech transcription systems improve continuously, leveraging multiple acoustic and language models similar to the techniques used here, as well as with innovations in modeling. Just since the submission of this article, results that improve on the 5.8% machine error rate for the Switchboard test set have been reported in [48] and [81], and our own system has been updated to achieve an error rate of 5.1% [46]. Further advances will no doubt be forthcoming.

X. CONCLUSIONS

We have conducted a two-pronged experiment comparing state-of-the-art speech recognition and human transcription for conversational telephone speech, as represented by the NIST evaluation set. First, we measured human accuracy on this task by submitting the data to a standard commercial transcription process, and postprocessed it for the NIST scoring protocol. We then assembled a high-accuracy recognition system based on a collection of neural-network-based acoustic and language models. Key features in achieving good performance included LSTMs for both acoustic and language modeling, together with CNNs in the acoustic model, as well as extensive

combination of complementary subsystems for both acoustic and language modeling. The recognition system achieved error rates of 5.8% and 11.0% on Switchboard and CallHome portions, respectively, slightly lower than corresponding human error.⁷ An analysis of differences in error patterns between the two version found remarkable similarity. The same speakers tend to be relative easy or hard to recognize for humans and machines, and the same kinds of short function words tend to be substituted, deleted or inserted in errors. Finally, human subjects had a hard time identifying which errorful transcripts originated from human or machines in a side-by-side comparison. For the Switchboard genre (conversations between strangers) our results support the conclusion that state-of-the-art speech recognition technology can reach a level that is comparable to human in both quantitative and qualitative terms, when given sufficient and matched training data.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for valuable feedback that greatly improved the paper. We are grateful to Arul Menezes for access to the Microsoft transcription pipeline; to Chris Basoglu, Amit Agarwal and Marko Radmilac for their invaluable assistance with CNTK; and to Jinyu Li and Partha Parthasarathy for many helpful conversations. We also thank X. Chen from Cambridge University for valuable assistance with the CUED-RNNLM toolkit, and the International Computer Science Institute for compute and data resources. Arlo Fario helped with metadata on the speakers in the evaluation data.

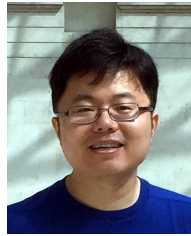
REFERENCES

- [1] M. Campbell, A. J. Hoane, and F.-h. Hsu, "Deep Blue", *Artificial intelligence*, vol. 134, pp. 57–83, 2002.
- [2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., "Mastering the game of Go with deep neural networks and tree search", *Nature*, vol. 529, pp. 484–489, 2016.
- [3] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos, et al., "Deep Speech 2: End-to-end speech recognition in English and Mandarin", arXiv preprint arXiv:1512.02595, 2015.
- [4] T. T. Kristjansson, J. R. Hershey, P. A. Olsen, S. J. Rennie, and R. A. Gopinath, "Super-human multi-talker speech recognition: the IBM 2006 Speech Separation Challenge system", in *Proc. Interspeech*, vol. 12, p. 155, 2006.
- [5] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Single-channel mixed speech recognition using deep neural networks", in *Proc. IEEE ICASSP*, pp. 5632–5636. IEEE, 2014.
- [6] D. S. Pallett, "A look at NIST's benchmark ASR tests: past, present, and future", in *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 483–488. IEEE, 2003.
- [7] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition", in *Proc. IEEE ICASSP*, pp. 651–654. IEEE, 1988.
- [8] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus", in *Proceedings of the workshop on Speech and Natural Language*, pp. 357–362. Association for Computational Linguistics, 1992.

⁷Arguably, based on the experiment in [48], human accuracy on CallHome can be substantially improved with more effort and contextual information. On the other hand, machine performance on CallHome is also severely hampered, namely, by lack of sufficient matched training data.

- [9] D. Graff, Z. Wu, R. MacIntyre, and M. Liberman, "The 1996 broadcast news speech and language-model corpus", in *Proceedings of the DARPA Workshop on Spoken Language Technology*, pp. 11–14, 1997.
- [10] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development", in *Proc. IEEE ICASSP*, vol. 1, pp. 517–520. IEEE, 1992.
- [11] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text", in *Proc. International Conference on Language Resources and Evaluation*, vol. 4, pp. 69–71, 2004.
- [12] S. F. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Soltau, and G. Zweig, "Advances in speech transcription at IBM under the DARPA EARS program", *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 1596–1608, 2006.
- [13] S. Matsoukas, J.-L. Gauvain, G. Adda, T. Colthurst, C.-L. Kao, O. Kimball, L. Lamel, F. Lefevre, J. Z. Ma, J. Makhouf, et al., "Advances in transcription of broadcast news and conversational telephone speech within the combined ears bbn/limsi system", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1541–1556, 2006.
- [14] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, et al., "Recent innovations in speech-to-text transcription at SRI-ICSI-UW", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1729–1744, 2006.
- [15] A. Ljolje, "The AT&T 2001 LVCSR system", NIST LVCSR Workshop, 2001.
- [16] J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefevre, "Conversational telephone speech recognition", in *Proc. IEEE ICASSP*, vol. 1, pp. I–212. IEEE, 2003.
- [17] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang, and P. C. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system", in *Proc. IEEE ICASSP*, vol. 1, pp. I–249. IEEE, 2004.
- [18] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks", in *Proc. Interspeech*, pp. 437–440, 2011.
- [19] R. P. Lippmann, "Speech recognition by machines and humans", *Speech Communication*, vol. 22, pp. 1–15, 1997.
- [20] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "The Microsoft 2016 conversational speech recognition system", in *Proc. IEEE ICASSP*, pp. 5255–5259, New Orleans, Mar. 2017.
- [21] F. J. Pineda, "Generalization of back-propagation to recurrent neural networks", *Physical Review Letters*, vol. 59, pp. 2229, 1987.
- [22] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks", *Neural Computation*, vol. 1, pp. 270–280, 1989.
- [23] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks", *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, pp. 328–339, 1989.
- [24] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series", *The handbook of brain theory and neural networks*, vol. 3361, pp. 1995, 1995.
- [25] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition", *Neural computation*, vol. 1, pp. 541–551, 1989.
- [26] T. Robinson and F. Fallside, "A recurrent error propagation network speech recognition system", *Computer Speech & Language*, vol. 5, pp. 259–274, 1991.
- [27] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
- [28] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling", in *Proc. Interspeech*, pp. 338–342, 2014.
- [29] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition", in *Proc. Interspeech*, pp. 1468–1472, 2015.
- [30] G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny, "The IBM 2015 English conversational telephone speech recognition system", in *Proc. Interspeech*, pp. 3140–3144, 2015.
- [31] T. Sercu, C. Puhersch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for LVCSR", in *Proc. IEEE ICASSP*, pp. 4955–4959. IEEE, 2016.
- [32] M. Bi, Y. Qian, and K. Yu, "Very deep convolutional neural networks for LVCSR", in *Proc. Interspeech*, pp. 3259–3263, 2015.
- [33] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition", *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, pp. 2263–2276, Aug. 2016.
- [34] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model", in *Proc. Interspeech*, pp. 1045–1048, 2010.
- [35] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model", in *Proc. Interspeech*, pp. 901–904, 2012.
- [36] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling", in *Proc. Interspeech*, pp. 194–197, 2012.
- [37] I. Medennikov, A. Prudnikov, and A. Zatzvornitskiy, "Improving English conversational telephone speech recognition", in *Proc. Interspeech*, pp. 2–6, 2016.
- [38] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models", in *Studies in Fuzziness and Soft Computing*, vol. 194, pp. 137–186, 2006.
- [39] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations", in *HLT-NAACL*, vol. 13, pp. 746–751, 2013.
- [40] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)", in *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 347–352, Santa Barbara, CA, 1997.
- [41] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation, and system combination", in *Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.
- [42] A. Stolcke et al., "The SRI March 2000 Hub-5 conversational speech transcription system", in *Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.
- [43] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks", in *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.
- [44] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al., "Deep speech: Scaling up end-to-end speech recognition", arXiv preprint arXiv:1412.5567, 2014.
- [45] A. Stolcke and J. Droppo, "Comparing human and machine errors in conversational speech transcription", in *Proc. Interspeech*, pp. 137–141, Stockholm, Aug. 2017.
- [46] W. Xiong, L. Wu, F. Allewa, J. Droppo, X. Huang, and A. Stolcke, "The Microsoft 2017 conversational speech recognition system", Technical Report MSR-TR-2017-39, Microsoft Research, Aug. 2017, <http://arxiv.org/abs/1708.06073>.
- [47] M. L. Glenn, S. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, "Transcription methods for consistency, volume and efficiency", in *Proc. International Conference on Language Resources and Evaluation*, 2010.
- [48] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, "English conversational telephone speech recognition by humans and machines", in *Proc. Interspeech*, pp. 132–136, Stockholm, Aug. 2017.
- [49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.
- [50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- [51] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks", *CoRR*, vol. abs/1505.00387, 2015.
- [52] P. Ghahremani, J. Droppo, and M. L. Seltzer, "Linearly augmented deep neural network", in *Proc. IEEE ICASSP*, pp. 5085–5089. IEEE, 2016.
- [53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", *Proceedings of Machine Learning Research*, vol. 37, pp. 448–456, 2015.
- [54] D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig, "Deep convolutional neural networks with layer-wise context expansion and attention", in *Proc. Interspeech*, pp. 17–21, 2016.
- [55] A. Waibel, H. Sawai, and K. Shikano, "Consonant recognition by modular construction of large phonemic time-delay neural networks", in *Proc. IEEE ICASSP*, pp. 112–115. IEEE, 1989.
- [56] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", *Neural Networks*, vol. 18, pp. 602–610, 2005.
- [57] N. Mesgaran, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus", *Science*, vol. 343, pp. 10061010, Feb. 2014.
- [58] C. Wu, P. Karanasou, M. J. Gales, and K. C. Sim, "Stimulated deep neural network for speech recognition", in *Proc. Interspeech*, pp. 400–404, 2016.
- [59] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification", *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.

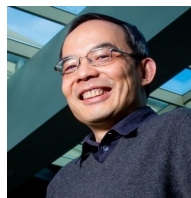
- [60] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors", in *IEEE Speech Recognition and Understanding Workshop*, pp. 55–59, 2013.
- [61] G. Saon, T. Sercu, S. J. Rennie, and H. J. Kuo, "The IBM 2016 English conversational telephone speech recognition system", in *Proc. Interspeech*, pp. 7–11, Sep. 2016.
- [62] G. Wang and K. Sim, "Sequential classification criteria for NNs in automatic speech recognition", in *Proc. Interspeech*, pp. 441–444, 2011.
- [63] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks", in *Proc. Interspeech*, pp. 2345–2349, 2013.
- [64] D. Povey, V. Peddinti, D. Galvez, P. Ghahmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI", in *Proc. Interspeech*, pp. 2751–2755, 2016.
- [65] C. Mendis, J. Droppo, S. Maleki, M. Musuvathi, T. Mytkowicz, and G. Zweig, "Parallelizing WFST speech decoders", in *Proc. IEEE ICASSP*, pp. 5325–5329. IEEE, 2016.
- [66] A. Stolcke, "SRILM—an extensible language modeling toolkit", in *Proc. Interspeech*, vol. 2002, p. 2002, 2002.
- [67] T. Alumäe and M. Kurimo, "Efficient estimation of maximum entropy language models with N-gram features: An SRILM extension", in *Proc. Interspeech*, pp. 1820–1823, 2012.
- [68] X. Chen, X. Liu, Y. Qian, M. J. F. Gales, and P. C. Woodland, "CUED-RNNLM: An open-source toolkit for efficient training and evaluation of recurrent neural network language models", in *Proc. IEEE ICASSP*, pp. 6000–6004. IEEE, 2016.
- [69] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models", *AISTATS*, vol. 1, pp. 6, 2010.
- [70] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data", in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pp. 2333–2338. ACM, 2013.
- [71] O. Press and L. Wolf, "Using the output embedding to improve language models", in M. Lapata, P. Blunsom, and A. Koller, editors, *Proc. 15th Conf. European Chapter of the Association for Computational Linguistics*, vol. 2, Short Papers, pp. 157–163, Valencia, Spain, Apr. 2017.
- [72] P. Ghahremani and J. Droppo, "Self-stabilized deep neural network", in *Proc. IEEE ICASSP*. IEEE, 2016.
- [73] M. Sundermeyer, H. Ney, and R. Schlüter, "From feedforward to recurrent LSTM neural networks for language modeling", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 517–529, Mar. 2015.
- [74] Microsoft Research, "The Microsoft Cognition Toolkit (CNTK)", <http://cntk.ai>.
- [75] D. Yu et al., "An introduction to computational networks and the Computational Network Toolkit", Technical Report MSR-TR-2014-112, Microsoft Research, 2014, <https://github.com/Microsoft/CNTK>.
- [76] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs", in *Proc. Interspeech*, pp. 1058–1062, 2014.
- [77] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", Proceedings 3rd International Conference for Learning Representations, arXiv preprint arXiv:1703.02136, May 2015.
- [78] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus", in *Proc. ICSLP*, 1996.
- [79] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR", in *Proc. IEEE ICASSP*, pp. 8614–8618. IEEE, 2013.
- [80] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition", in *Proc. IEEE ICASSP*, pp. 4277–4280. IEEE, 2012.
- [81] K. J. Han, S. Hahm, B.-H. Kim, J. Kim, and I. Lane, "Deep learning-based telephony speech recognition in the wild", in *Proc. Interspeech*, pp. 1323–1327, Stockholm, Aug. 2017.



Wayne Xiong (M'07) received the Bachelor degree in electrical engineering from Tsinghua University, Beijing, P. R. China, in 2005, and the Master degree in computer science and engineering from the Chinese University of Hong Kong, Hong Kong, in 2007. He joined Microsoft Corporation in 2007. He is currently working as a Senior Researcher in the Speech and Dialog Research Group. His research interests include speech recognition and next generation of dialog systems.



Jasha Droppo (M'03-SM'07) received the B.S. degree in electrical engineering (with honors) from Gonzaga University, Spokane, WA, in 1994, and the M.S. and Ph.D. degrees in electrical engineering from the University of Washington, Seattle, in 1996 and 2000, respectively. At the University of Washington, he helped to develop and promote a discrete theory for time-frequency representations of audio signals, with a focus on speech recognition. He is best known for his research in robust speech recognition, including algorithms for speech signal enhancement, model-based speech feature enhancement, robust speech features, model-based adaptation, and noise tracking. His current interests include the use of neural networks in acoustic modeling and the application of large data and general machine learning algorithms to previously hand-authored speech recognition components.



Xuedong Huang (M'89-SM'94-F'00) is a Microsoft Technical Fellow in AI and Research. He leads Microsoft's Speech and Language Group, having joined Microsoft to found the company's speech technology group in 1993. Dr. Huang received his Ph.D., M.S., and B.S. degrees from the University of Edinburgh, Tsinghua University, and Hunan University, respectively, and has previously worked at Carnegie Mellon University (1989-1993). He has been recognized with the Alan Newell Research Excellence Leadership Medal in 1992, with an IEEE Paper Award in 1993, as the Asian American Engineer of the Year in 2011, and as one of Wired Magazine's 25 Geniuses in 2016. Dr. Huang holds over 100 patents and has published over 100 papers.



Frank Seide is a Principal Researcher at Microsoft's AI and Research division, and an architect of Microsoft's Cognitive Toolkit for deep learning. His research focus is on deep neural networks for speech recognition and machine translation. Together with co-author Dong Yu, he was first to show the effectiveness of deep neural networks for recognition of free conversations. Since graduation in 1993, Frank Seide has worked on various speech and language topics, first at Philips Research in Aachen and Taipei, then at Microsoft Research Asia (Beijing), and now at Microsoft Research (Redmond), with contributions to spoken dialog systems, Mandarin speech recognition, audio search, speech-to-speech and text translation, and distributed neural network model training.



Mike Seltzer (M'95-SM'07) received the Sc.B. degree with honors from Brown University in 1996 and M.S. and Ph.D. degrees from Carnegie Mellon University in 2000 and 2003, respectively, all in electrical engineering. From 1998 to 2003, he was a member of the Robust Speech Recognition Group at Carnegie Mellon University. Since 2003, he has been a member of the Speech and Dialog Research Group at Microsoft Research, where he is currently a Principal Researcher. In 2006, he was awarded the IEEE SPS Best Young Author paper award for

his work on microphone array processing for speech recognition. Dr. Seltzer has made scientific contributions in noise robustness, acoustic modeling and speech enhancement, and he has contributed to several Microsoft products including Microsoft Cognitive Services, Microsoft Cognitive Toolkit, Cortana, Bing Voice Search, Windows Phone, and Windows Automotive. He is currently a member of the IEEE Signal Processing Society Speech and Language Technical Committee (SLTC). From 2006-2008, he was Editor-in-Chief of the SLTC e-Newsletter and from 2009-2011, he was an Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing. His current interests include speech recognition in adverse environments, acoustic modeling and adaptation, neural networks, microphone arrays, and machine learning for speech and audio applications.



Geoff Zweig (M'98-SM'07-F'13) received his Ph.D. in Computer Science, with a major in Artificial Intelligence, from the University of California at Berkeley in 1998. From 1998 to 2006, Dr. Zweig was a Research Staff Member at IBM Research, where his work focused on new speech recognition technologies. From 2003 through 2006, he was the Manager of the Advanced Large Vocabulary Speech Recognition Research Group at IBM, where he was responsible for developing English, Arabic and Mandarin speech recognition systems for the DARPA

EARS and GALE programs. From 2006 through 2016, Dr. Zweig worked at Microsoft Research, most recently as a Partner Research Manager, in charge of the Speech and Dialog Research Group. At Microsoft, his research included work on speech recognition, language modeling, dialog systems, and image captioning, and found application in Microsoft's Cognitive Services. He is currently the Head of Machine Learning at a large financial company.

Dr. Zweig has published over 80 papers, served as an Associate Editor of the IEEE-TASLP, and of *Computer Speech & Language*, served on the SLTC of the IEEE Signal Processing Society.



Andreas Stolcke (M'95-SM'05-F'11) received a Diplom in computer science from Technical University Munich in 1988, and the Ph.D. degree in computer science from the University of California, Berkeley, in 1994. Until 2011 he was a Senior Research Engineer in the Speech Technology and Research Laboratory at SRI International, Menlo Park, CA, and is currently a Principal Researcher with the Speech and Dialog Research Group at Microsoft Research in Sunnyvale, CA; he is also an External Fellow at the International Computer

Science Institute in Berkeley, CA.

Dr. Stolcke has published over 200 papers in machine language learning, parsing, language modeling, speech recognition, speaker recognition, and speech understanding, and is the author of the widely used SRILM open-source toolkit for statistical language modeling. He has served on the editorial board of *Computational Linguistics*, as an Associate Editor of the IEEE-TASLP, and as a co-editor of *Computer Speech & Language*, and is currently on the SLTC of the IEEE Signal Processing Society.



Dong Yu (M'97-SM'06) is a distinguished scientist and vice general manager at Tencent AI Lab. Before joining Tencent, he was a Principal Researcher at Microsoft Research, which he had joined in 1998. His research on deep learning based speech recognition has been recognized by the IEEE Signal Processing Society 2013 and 2016 Best Paper award. He has served on several technical committees, editorial boards, and conference organization committees.