HOW NEURAL NETWORK FEATURES AND DEPTH MODIFY STATISTICAL PROPERTIES OF HMM ACOUSTIC MODELS

Suman Ravuri[1,2]*, Steven Wegmann*[1,3]

[1]International Computer Science Institute,
[2]University of California, Berkeley, CA, USA,
[3]Semantic Machines, Inc.

ravuri@icsi.berkeley.edu, swegmann@icsi.berkeley.edu

ABSTRACT

Tandem neural network features, especially ones trained with more than one hidden layer, have improved word recognition performance, but why these features improve automatic speech recognition systems is not completely understood. In this work, we study how neural network features cope with the mismatch between the underlying stochastic process inherent in speech, and the models we use to represent that process. We use a novel resampling framework, which resamples test set data to match the conditional independence assumptions of the acoustic model, and measure performance as we break those assumptions. We discover that depth provides modest robustness to data/model mismatch at the state level, and compared to standard MFCC features, neural network features actually fix poor duration modeling assumptions of the HMM. The duration modeling problem is also fixed by the language model, suggesting that the dictionary and language model make very strong implicit assumptions about phone length, which may now need to be revisited.

Index Terms— Neural Networks, Deep Learning, Tandem Features, Hidden Markov Models

1. INTRODUCTION

The recrudescence of neural networks as a research focus for Automatic Speech Recognition (ASR) systems emerges from a growing body of empirical evidence showing that the use of such models improves word recognition performance. Since work in [1], modifications to neural network models have led to a steady drop in recognition error rates, and perhaps a bit unsurprisingly, more research has focused on exploring new models rather than trying to understand what exactly is causing improvements in existing ones.

Despite the rapid evolution in neural network models for ASR – proposed models such as CTC-trained RNNs [2] attempt to replace the now-standard DNN-HMM acoustic model–, one element has remained constant across a variety of systems: a frame-level classification with a neural network using multiple hidden layers. One subclass of models is the so-called "hybrid" system, in which the GMM frame classifier of the GMM-HMM system is replaced with a DNN. Within this subclass, there have notable attempts in understanding how these systems improve recognition. [3] compared DNN-HMM to GMM-HMM systems by comparing DNN models to GMMs on phone error rate, noise robustness, and speaking rate, and concluded that DNNs are likely better frame estimators than GMM. A separate attempt – [4] – measured the ASR performance after each step of MFCC processing. Recently, [5] showed that hidden units of deeper layers encoded more specific phonemic information, while also stripping away seemingly uninformative properties such as gender. For deep Tandem features, there was an early attempt at comparing depth in [6], comparing frame error rates of a three-hidden-layer MLP to one with a single layer and its effect on word error rate in noise-added conditions.

This work adds to this body of literature by trying to understand if and how neural networks modify the statistical properties of the models we use for Automatic Speech Recognition. In particular, given that we assume speech to be a stochastic process with distribution $\mathbb{P}_{true}(O, W)$, and we represent this random process with a model with distribution $\mathbb{P}_{model}(O, W) \approx \mathbb{P}_{model}(O|W)\mathbb{P}_{model}(W)$, a natural question to ask is how well do our models match the true distribution. We would like to understand how the model mismatch – i.e., the difference between $\mathbb{P}_{model}$ and $\mathbb{P}_{true}$ – affects ASR performance. Unfortunately, direct access to $\mathbb{P}_{true}(O, W)$ is difficult, so we instead construct synthetic data to match the conditional independence assumptions of our models, and measure performance as we break those conditional independence assumptions.

We use the resampling process described in [7, 8, 9], which uses simulation and novel sampling process to generate pseudo test data that deviate from the HMM in a controlled fashion. These processes allow us to generate pseudo data that, at one extreme, agree with all of the model's assump-

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1450916.

tions, and at the another extreme, deviate from the model in exactly the way real data do. In between, we can precisely control the degree of data/model mismatch. By measuring recognition performance on this pseudo test data, we are able to quantify the effect of this controlled data/model residual on recognition accuracy. The novel sampling process, called resampling, was adapted from Bradley Efron's work on the bootstrap [10, 11]. Segment level resampling creates pseudo test data by randomly sampling (with replacement) labeled—using forced alignment—segments from real test data; the resulting pseudo test data is independent between the segments and inherits whatever dependence present in the segments. In this paper we use frame-, state-, and phone-level resampling. See [7] for a detailed description.

In this work, we explore how Tandem features and depth of the neural networks used to generate those features affect the statistical properties of ASR models. While it may seem a bit parachronistic to explore Tandem features given the near ubiquitous use of hybrid systems, the similarities between the two systems – both generate "derived" features based on supervised training on phone-like alignments – may provide insights into both Tandem and hybrid systems. Moreover, "deep" Tandem features have been known to improve recognition performance even in hybrid systems (such as in low resource settings [12]), so they are worth studying in their own right.

One additional issue is that since ASR systems are rather complex, it is not at all obvious how changes to a particular feature affect downstream processing. Even for a basic ASR system, a Deep Neural Network (DNN) or Gaussian Mixture Model (GMM) frame classifier estimates the a (pseudo-)likelihood of a context-dependent triphone for a particular feature, a temporal model such as a hidden Markov Model (HMM) generates phone sequence estimates from frame likelihoods, a lexicon restricts allowable phone sequences to those consistent with actual words, and a language model provides likelihood estimates for sequences of words. Problems fixed at a feature level may already be fixed later, or may break a hack used in another part of the system. As a result, we perform our analysis with just the HMM as a phone loop without a language model or lexicon constraints, and then for the entire recognition pipeline. For the HMM phone loop, we calculate phoneme error rates and marginal phone duration lengths of the predictions as the test set data moves from matching the conditional independence assumptions of the model to more realistic test data. Then we redo this analysis when we include the language model and lexicon, and also include word error rate results.

In this work, we would like to answer four questions: 1) are neural network features more robust to model mismatch than more standard frame-level features, 2) does depth provide more robustness to data/model mismatch, 3) how does the choice of feature affect expected phone duration of predictions using only the HMM phone loop, and 4) does including language model information change the predicted phone duration length? In brief, the experiments suggest that using neural network features are quite a bit more robust to data/model mismatch than MFCCs, and depth provides additional robustness to state- and phone-level statistical dependence. Moreover, using neural network features fix poor duration modeling assumptions of the acoustic model. The language model fixes also phone duration modeling, while the MLP features has this benefit without needing to include lexicon and language model constraints.

2. EXPERIMENTAL SETUP

2.1. Data and Modeling

We use the spontaneous meeting portion of the ICSI meeting corpus [13], recorded with near-field microphones. The training set consists of 23,739 utterances – 20.4 hours – of speech across 26 speakers. The training set is based on meeting data used for adaptation in the SRI-ICSI meeting recognizer [14]. Since phone loop recognition is quite fast, the test set is a disjoint 20 hour set from the ICSI meeting corpus. Since decoding using the full recognition system is much slower, we use a test set comprising 58 minutes of speech, taken from ICSI meetings portions of the NIST Rich Transcription Evaluation Sets 2002 [15], 2004 [16], and 2005 [17]. Resampling for this latter test set is performed 5 times to determine the variance in sampling. Previous work [8, 18, 9] use this setup with an HTK recognizer, and a more complete description of the setup can be found in [8].

The acoustic models use cross-word triphones and are estimated using maximum likelihood. Each triphone is a three-state linear HMM with no skipping, except for the silence phone. The output distribution is a single Gaussian, since we are not necessarily interested in the best results but merely those for analysis. Maximum likelihood training roughly follows the HTK tutorial: monophone models are estimated from a "flat start", duplicated to form triphone models, clustered to 2,500 states and re-estimated. We use HDecode for decoding with a wide search beam (300) to avoid search errors. To evaluate recognition accuracy the reference and the decoded utterances are text normalized before the NIST tool sclite is used to obtain word error rate (WER).

We use a trigram language model (LM) [14] that was trained at SRI by interpolating a number of source LMs; these consisted of webtext and the transcripts of the following corpora: Switchboard, meetings (CMU, ICSI, and NIST), Fisher, Hub4-LM96, and TDT4. We renormalized the language model after removing words not present in the training dictionary. The perplexity of this meeting room LM is around 70 on our test set. To be compatible with the SRI LM, we use the SRI pronunciation dictionary, which includes two extra phones compared to the CMU phone set – "puh" and "pum" – to model hesitations.

2.2. Features

In this work, we compare MFCC to Tandem [19] features. The MFCCs are generated by the HTK Front-End, with 13 Mel-cepstral coefficients, including energy, and first and second differences. 9 Frames of MFCC features serve as input to the neural network, which is trained using TNet [20]. The number of hidden layers for this study range from 1-4, and we found severely degraded performance using more than 4 hidden layers. Each layer consisted of 1,500 hidden units, as this produced the best results in initial experiments, and each hidden unit used a sigmoid non-linearity. The networks were pretrained [21] before cross-entropy training. The labels were 42 phone targets, generated from alignments using a GMM-HMM baseline system with 2,500 states and 8 Gaussians per mixture. Training converged for all neural networks after 13-15 epochs.

2.3. Metrics and Alignments

Our study tracks three metrics: phone error rate, word error rate, and phone duration. The reference phone lengths depend on alignments, which are generated using the same alignments used for training of the neural networks. Since alignments using different features may differ by 10% [22], and using the same alignment for both training and test may bias results in favor of MLP-based features, we also performed a preliminary study on alignment agreement and calculated frame error using alignments generated from a 8 Gaussian mixture GMM-HMM system using 3 hidden layer MLP features. While we found alignment disagreement to be around 5%, the relative ordering of performance of features did not change, so for this work we report on only reference phone lengths generated from MFCC alignments.

In addition to the above caveats, phone durations are generated from state-level alignments, which are subject to misalignment. In particular, alignments that are unable to locate particular phones will default to the minimum duration of three frames, due to structural constraints of the three-state Bakis phone hidden Markov Models. Figures 1 and 2 show a large percentage of phones with a duration of 3 frames in frame-level resampled and original test data, but this mode is more likely due to alignment error than another effect.

For resampling experiments, extra care must be taken as lengths of utterances change at the state-, phone-, and word-level. After test utterances are regenerated under the sampling framework, we realign the sampled data using an 8 Gaussian mixture GMM-HMM system with MFCC features and use those alignments as a reference. We also compared against alignments using MLP-based features, but found no significant difference in results.

3. RESULTS

3.1. HMM Phone Loop

Table 1 shows the phone error rate for MFCC features and neural network features by depth. When the conditional independence assumptions are matched at the frame level, there is little benefit replacing MFCCs with Tandem features; in fact, MFCC features outperform MLP based features in all but the three hidden layer case. Starting with the state-level resampling, however, neural network features significantly outperform MFCC features. While the phone error rate degrades as the data becomes more realistic for all features, MLP-based features degrade less rapidly.

To understand why there is a significant difference between frame- and state-level results between the two types of features, it is instructive to look at Figure 1. At the frame-level, the expected duration of MFCC features match the durations from the alignment (including the spurious peak at the minimum phone duration length), but at more realistic sampling levels, the percentage of phones of duration 8 frames or longer is severely underestimated. In contrast, neural network features match the longer phone duration lengths more accurately. In some sense, MLP-based features are actually fixing the poor duration modeling assumptions of phone HMMs.

Among MLP-based classifiers, using two hidden layers instead of one seems to provide some modest robustness to data/model mismatch when moving from frame- to state-level resampling (shown in the bottom portion of Table 1). At more realistic sampling levels, however, relative degradation seems to be flat.

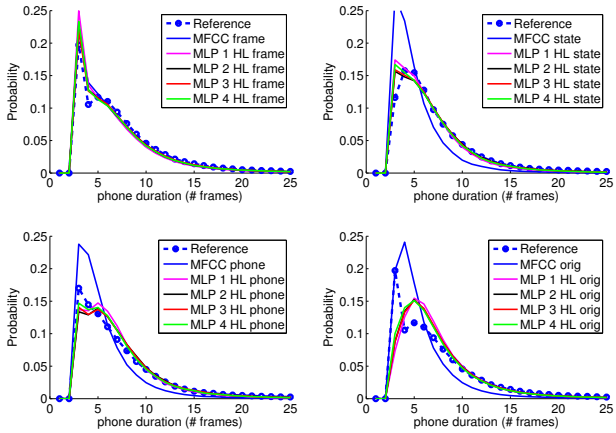

Fig. 1. *Reference vs. Model Phone Duration Histograms for different features and resampling units using an HMM Phone loop. Word-level resampling results are omitted due to space constraints, but are similar to phone-level results.*

3.2. Full Recognition System with Lexicon and Language Model

As expected, including language model information substantially improves phone error rate results for all features, as

	MFCC	MLP 1HL	MLP 2HL	MLP 3HL	MLP 4HL
frame	15.77	19.74	15.94	13.73	15.88
state	83.83	42.58	33.18	34.34	35.36
phone	86.74	51.60	43.47	42.42	42.11
word	91.43	60.13	54.24	51.30	53.64
original	93.10	61.64	59.56	58.39	58.95
frame/state	431%	116%	108%	150%	122%
state/phone	3.47%	21.2%	31.0%	23.5%	19.1%

Table 1. *Phone Error Rate for HMM Phone Loop for different types of resampled data (top), and relative degradation among different types of features (bottom).*

shown in Table 2, and better phone recognition correlates well – but not perfectly – with word recognition results, shown in Table 3. The result is not terribly surprising, as the lexicon restricts allowable phone sequences to correspond to actual words. Moreover, compared to results using only the HMM phone loop, the system is also more robust to conditional independence assumption mismatches across all features, and especially for MFCCs. Figure 2 shows one possible cause: the underestimates of phones 8 frames or higher has now vanished. Including a language model seems to fix the poor duration model of the HMM phone model.

Even though the LM does fix data/model mismatch problems, MLP features are still more robust than MFCCs, especially at the state level (shown in the bottom part of Table 2). Moreover, depth up to three hidden layers seems to improve this type of robustness to statistical dependence at the state level. At other sampling levels, relative degradation seems to be static.

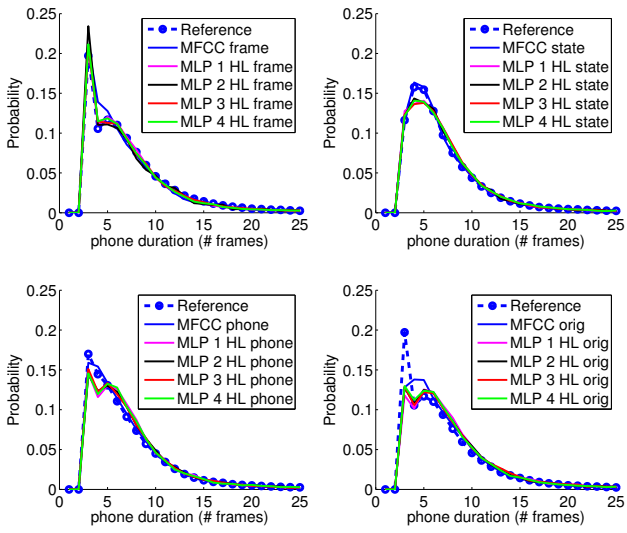

Fig. 2. *Reference vs. Model Phone Duration Histograms for different features and resampling units using the full recognition system. Word-level resampling results are omitted due to space constraints, but are similar to phone-level results.*

	MFCC	MLP 1HL	MLP 2HL	MLP 3HL	MLP 4HL
frame	4.52(.03)	3.02(.01)	2.57(.08)	2.11(.02)	2.28(.02)
state	8.37(.19)	5.02(.10)	3.80(.13)	2.47(.24)	4.20(.08)
phone	15.8(.18)	10.1(.28)	7.00(.16)	5.03(.11)	6.52(.20)
word	30.8(.43)	25.5(.64)	19.6(1.2)	18.9(.35)	19.3(.47)
orig	32.08	27.56	22.57	21.04	21.55
frame/state	85.2%	66.2%	47.9%	17.1%	84.2%
state/phone	88.8%	101%	84.2%	103%	55.2%

Table 2. *Phone Error Rate using full recognition system for different types of resampled data (top), and relative degradation between different types of features (bottom). Numbers in parentheses refer to standard deviation of error across 5 runs of resampled data.*

	MFCC	MLP 1HL	MLP 2HL	MLP 3HL	MLP 4HL
frame	1.02(.11)	0.78(.05)	0.86(.05)	0.70(0.0)	0.80(0.0)
state	7.30(.23)	4.46(.21)	3.40(.19)	4.48(.29)	3.92(.18)
phone	20.6(.42)	13.6(.34)	9.52(.19)	9.48(.36)	8.82(.46)
word	37.3(.82)	31.3(.68)	25.4(0.26)	21.9(.11)	21.8(.62)
orig	44.6	40.0	32.6	31.6	31.8
frame/state	616%	472%	295%	540%	390%
state/phone	182%	205%	180%	111%	125%

Table 3. *Word Error Rate using full recognition system for different types of resampled data (top), and relative degradation between different types of features (bottom). Numbers in parentheses refer to standard deviation of error across 5 runs of resampled data.*

4. CONCLUSION

In this work, we track how neural network features improve ASR systems by testing performance on phone classification and phone duration modeling in two settings: using only a HMM phone loop, and full recognition system. Moreover, we compared how neural network features coped with data/model mismatch by comparing recognition performance on the original test data to that of data resampled to better match the model's conditional independence assumptions. We found that depth improves data/model mismatch robustness at the state level using the full recognition system. Moreover, neural network features themselves fix poor phone duration modeling assumptions of the hidden Markov Model. These poor modeling assumptions, though, are already fixed by including the dictionary and language model.

Prima facie, it seems as if duration modeling should be handled by the HMM phone model, or barring that, the acoustic model. That the lexicon and LM, which in and of itself do not explicitly model phone duration, also fix phone durations seems especially troubling. We encounter these problems when we tune recognizers: we scale language model scores to account for, among other things, score mismatch between the acoustic and language models, only to include a separate word insertion penalty, because increasing language model scaling factors now results in hypothesized word sequences with fewer longer words. That neural network features "fix" the phone duration model suggest that other improvements in the model remain.

5. REFERENCES

[1] G.E. Dahl, Dong Yu, Li Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30 –42, jan. 2012.

[2] Andrew L. Maas, Ziang Xie, Dan Jurafsky, and Andrew Y. Ng, "Lexicon-free conversational speech recognition with neural networks," in *North American Chapter of the Association for Computational Linguistics*, Singapore, 2015.

[3] Yan Huang, Dong Yu, Chaojun Liu, and Yifan Gong, "A comparative analytic study on the gaussian mixture and context dependent deep neural network hidden markov models," in *Interspeech 2014*, September 2014.

[4] Zoltán Tüske, Pavel Golik, Ralf Schlüter, and Hermann Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr," in *Interspeech*, Singapore, Sept. 2014, pp. 890–894, ISCA best student paper award Interspeech 2014.

[5] Tasha Nagamine, Mike Seltzer, and Nima Mesgerani, "Exploring how deep neural networks form phonemic categories," in *Proc. Interspeech*, Dresden, Germany, 2015.

[6] O. Vinyals and S.V. Ravuri, "Comparing multilayer perceptron to deep belief network tandem features for robust asr," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4596–4599.

[7] Dan Gillick, Larry Gillick, and Steven Wegmann, "Don't multiply lightly: Quantifying problems with the acoustic model assumptions in speech recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11-15, 2011*, 2011, pp. 71–76.

[8] Sree Hari Krishnan Parthasarathi, Shuo-Yiin Chang, Jordan Cohen, Nelson Morgan, and Steven Wegmann, "The blame game in meeting room asr: An analysis of feature versus model errors in noisy and mismatched conditions," in *ICASSP'13*, 2013, pp. 6758–6762.

[9] S.-Y. Chang and Steven Wegmann, "On the importance of modeling and robustness for deep neural network feature," in *Proc. ICASSP*, April 2015.

[10] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, vol. 7, pp. 1–26, 1979.

[11] Bradley Efron, *The Jackknife, the bootstrap and other resampling plans*, CBMS-NSF Reg. Conf. Ser. Appl. Math. SIAM, Philadelphia, PA, 1982, Lectures given at Bowling Green State Univ., June 1980.

[12] Shakti P. Rath, Kate M. Knill, Anton Ragni, and Mark J. F. Gales, "Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages," in *Proc. Interspeech*, Singapore, 2014.

[13] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters, "The icsi meeting corpus," in *Proc. Interspeech*, 2003, pp. 364–367.

[14] Oliver Cetin and Andreas Stolcke, "Language modeling in the icsi-sri spring 2005 meeting speech recognition evaluation system," Tech. Rep., International Computer Science Institute, 2005.

[15] "Rt-2002 evaluation plan, `http://www.itl.nist.gov/iad/mig/tests/rt/2002/docs/rt02_eval_plan_v3.pdf.`," .

[16] "Rt-04s evaluation data documentation, `http://www.itl.nist.gov/iad/mig/tests/rt/2004-spring/eval/docs.html`," .

[17] "Rt-05s evaluation data documentation, `http://www.itl.nist.gov/iad/mig/tests/rt/2005-spring/eval/docs.html`," .

[18] Suman V. Ravuri, "Hybrid mlp/structured-svm tandem systems for large vocabulary and robust ASR," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 2729–2733.

[19] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 3, pp. 1635–1638, 2000.

[20] Stanislav Kontár, "Parallel training of neural networks for speech recognition," in *Proc. 12th International Conference on Soft Computing MENDEL'06*. 2006, p. 6, Brno University of Technology.

[21] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[22] Andreas Stolcke, Neville Ryant, Vikramjit Mitra, Wen Wang, and Mark Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *Proc. IEEE ICASSP*, Florence, May 2014, pp. 5589–5593, IEEE SPS.