TOWARDS MACHINES THAT KNOW WHEN THEY DO NOT KNOW: SUMMARY OF WORK DONE AT 2014 FREDERICK JELINEK MEMORIAL WORKSHOP

Hynek Hermansky[1,2], Lukáš Burget[2], Jordan Cohen[3], Emmanuel Dupoux[4], Naomi Feldman[5], John Godfrey[5], Sanjeev Khudanpur[1], Matthew Maciejewski[6], Sri Harish Mallidi[5], Anjali Menon[6], Tetsuji Ogawa[7], Vijayaditya Peddinti[1], Richard Rose[8], Richard Stern[6], Matthew Wiesner[8], Karel Veselý[2]

[1]Johns Hopkins University, [2]Brno University of Technology, [3]Spelamode, [4]Ecole Normale Supérieure, [5]University of Maryland, [6]Carnegie Mellon University, [7]Waseda University, [8]McGill University

ABSTRACT

A group of junior and senior researchers gathered as a part of the 2014 Frederick Jelinek Memorial Workshop in Prague to address the problem of predicting the accuracy of a nonlinear Deep Neural Network probability estimator for unknown data in a different application domain from the domain in which the estimator was trained. The paper describes the problem and summarizes approaches that were taken by the group[1].

Index Terms— Performance monitoring, confidence estimation, multistream recognition of speech.

1. INTRODUCTION

An implicit assumption in machine learning is that the harmful variability that is encountered in use of a machine is drawn from the same distribution as the variability that was present in the training data. However, in practice, obtaining training data that covers all unexpected variability is difficult, if not impossible [1]. When faced with unexpected variability, a machine that has not been trained on this particular type of variability can generate spurious outputs that do not represent relevant aspects of the input. Thus, an estimator that is effective on data from one domain may perform poorly on data from another domain.

Identifying such domain anomalies [2] is desirable. Because the goal is to identify those mismatches that result in poor generalization to the new domain, it is not sufficient to simply identify data outliers. Instead, it is necessary to identify data that do not produce reliable results in a particular estimator.

A question then arises as to how one might determine the reliability of an estimator without already knowing the desired output of the computation. Evaluating the quality of the output requires predicting its accuracy in an unsupervised setting. In this paper we explore the possibility that, even when we do not know what the correct outputs for a given input should be, we sometimes know general characteristics that the output from the estimator should exhibit. These characteristics can often be learned from the estimator's performance on training data.

We focus on the specific case of a front-end speech recognition system that estimates the posterior probabilities of phonemes given the speech signal. The dominant technique here is currently the Artificial Neural Net (ANN) discriminative technique, and our focus was on this approach. Figure 1 shows posteriograms (estimates of posterior probabilities as a function of time) for phoneme labels, for posteriors derived from data that are similar to training data, and for posteriors derived in a domain that is different from the domain of the training data. It appears possible to estimate visually which posteriogram is more accurate. Our efforts examine ways of quantifying this intuition by machine.

The problem we address is:

Given an estimator that yields a vector of posterior probabilities of phonemes for every 10 ms of speech input, predict the accuracy of these estimates.

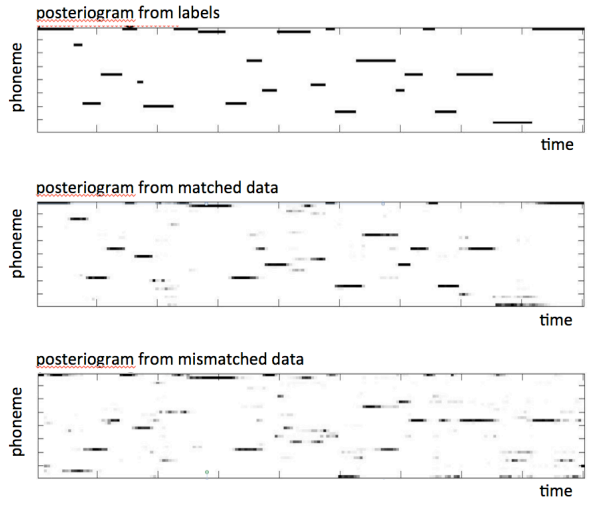

Fig. 1 Posteriograms from labels (top, matched data (middle), and mismatched data (bottom).

This work was supported in parts by the National Science Foundation via award number IIA-0530118, by grant from Google Inc., and by Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, IARPA, DoD/ARL, or the U.S. Government.

2. RELATED WORK

Prediction of accuracy of the estimation is related to the extensively studied task of estimation of confidence in ASR (see, e.g., [3] for the current state of the art). However, these tasks are not identical. In the confidence estimation we need to determine how reliable the result of the estimation is. In our case, we would like to know how accurate the result of the estimation is, without knowing what the true answer should be. Still, a number of techniques that originated in confidence estimation for GMM-based ASR can be also applied in predicting the accuracy of the estimation for ANN-based ASR. One frequently-used standard technique that we investigated in this work and that is applicable to ANN-based ASR requires a full Viterbi search for the best path through the probability estimates, is based on the averaged likelihood of the recognized sound sequence.

Several ANN-specific techniques were used for the prediction of the estimation accuracy from discriminative ANN estimators in the past. Among these are comparison of the highest-probability estimate to the next several lower ones [4] and a related technique based on the entropy of the estimator output [5][6]. Another previously proposed technique (not studied here) is based on the autocorrelation matrix of transformed probability estimates [7][8]. We adopted the entropy-based technique as a baseline technique in our evaluations in the workshop.

The technique that was the most effective prior to the workshop evaluates averaged dissimilarities of probability estimates spaced in several time spans apart (denoted here as the M-measure) [8][9]. This technique represents another baseline for our evaluations.

3. DATA USED DURING THE WORKSHOP

Data from 31 different probability estimators trained and tested using the TIMIT database (continuous read speech, adult males and females) were used during the workshop. In training the probability estimators we used about 200 minutes of the TIMIT data in the form of either a) the original (clean) speech data, or b) training speech data corrupted by various levels of subway noise.

A deep ANN spectral band probability estimator (4 hidden layers, 1000 nodes each), was trained on each of frequency band. Concatenated posteriors from each of 31 nonempty combinations of the spectral bands were used as inputs to 31 different merging probability estimators, also implemented as deep ANNs. Each of these merging estimators was trained on clean training data. In that way, 31 probability estimators were constructed, each stream using some combination of 5 spectral bands, from one spectral band up to all five spectral bands. The estimator using all five spectral bands is shown in Figure 2.

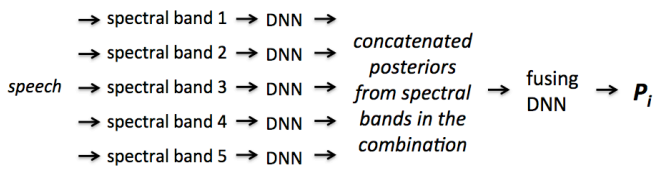

Fig. 2 Probability estimator using all five spectral bands.

All test data were processed by each of the 31estimators. Thus, altogether there were 310 different posterior streams to be evaluated, each containing data from 400 TIMIT sentences The test data for the probability estimations consisted of about 20 minutes (400 sentences) of the original test speech data and the test data corrupted by

- Nine various additive noises: Seven broadband noises from the Noisex database (clean speech, speech babble 15 dB SNR, car noise at 5 dB SNR, exhibition hall noise at 5 dB SNR, factory noise at10 dB SNR, restaurant noise at 10 dB SNR, street noise at 5 dB SNR, subway noise at15 dB SNR), and two narrow-band noises that fall in the 2nd and 4th frequency bands that were derived from the exhibition hall noise.

- Band-pass filtering the signal by forming various combinations of the 5 spectral bands.

Additionally, some experiments were performed on estimators that had been trained on data corrupted by different levels of the exhibition hall noise used different speech recognizer. These results are not reported or discussed in this paper.

4. EVALUATION CRITERIA

We applied the following evaluation criteria:

- Correlations with divergences from perfect probability values (Fig. 1 upper part - labels).
- Correlation with divergences from probability estimates derived on training-like test data (Fig. 1 middle part).
- Correlations with accuracies of a phoneme recognizer that uses the probability estimates.
- Phoneme recognition accuracy of a multistream adaptive ASR system that uses the estimates.

The first three criteria evaluate all predictions, the last one only evaluates whether the predictor can identify the best estimator.

5. OVERVIEW OF TECHNIQUES STUDIED

5.1. Delta M-measure

The original M-measure evaluates averaged divergences between estimates coming from different sounds. An extension of this technique, inspired by the segmentation algorithm proposed in [10], uses knowledge of average phoneme length derived from the training data, and computes the difference in divergences coming from the same sound and from different sounds. An interesting and successful extension computes the same/different sound probabilities in several time spans and solves the set of over-specified linear equations. This technique was studied the most extensively and has yielded the best results to date.

5.2. ANN-based autoencoder

An ANN is trained on the training data to predict estimate of posterior probabilities of phonemes from itself. To prevent a trivial solution, the ANN contains a bottleneck layer that is smaller than the output vector. This trained ANN is applied to the test data. The inverse of the error of the estimate is used as a predictor. This

technique, described in more detail elsewhere [11], appears to be promising.

5.3. Fit of unigram and bigram probabilities

Unigram and bigram probabilities of speech sounds estimated from labels (gold transcripts) or from posteriograms derived from training data (decoded transcripts) are compared to unigram and bigram probabilities computed from sequences of top probability estimates on the test data. This technique is intuitively appealing because the goodness of the estimate is judged by its degree of fit to expected linguistic information. Further work is required to fully understand its strengths and weaknesses.

5.4. Values of hidden units of an ANN

Several models of the distributions of values of the hidden units of the ANN are constructed on training data of the estimator. Outputs from these models are combined to predict the accuracy of the phoneme probability estimates. This technique is unique among those studied in that it looks inside the ANN for information about its performance on unknown test data, rather than looking only at the output. Details of this technique with results it yields are reported elsewhere [12].

5.6. Deviations from speech manifold acquired in the training of the estimator

This technique, based on a recent work described in [13] and assumes that speech lies on a low-dimensional manifold that can be learned during the training of the estimator. Significant deviations from this manifold that are detected during testing may indicate data-domain mismatch. More work is required to evaluate its full potential in application that were targeted in the workshop. .

6. MULTI-STREAM BASED ADAPTATION OF THE PROBABILITY ESTIMATOR

Multi-stream processing [14][15][16] is one biologically consistent way of adapting the classifier, which capitalizes on redundancies in coding of speech information. Different parallel processing streams attend to different aspects of redundantly coded information. When some streams are corrupted, the remaining uncorrupted streams can often still be used for decoding the message in the speech. This technique, unlike many others introduced for GMM-based estimators, is directly applicable to the currently-dominant ANN-based ASR.

Our probability estimators were constructed by exploiting various parts of the available speech spectrum. This allows for testing our techniques in this multi-stream adaptation paradigm. In this test, our techniques for predicting the estimator accuracy is used as a "performance monitor", i.e., we report the phoneme recognition accuracy from the phoneme recognizer (using Viterbi search on the estimated phoneme likelihoods using the bi-gram phonotactic language model), which uses the probability estimator that is predicted as the best one. Even though selecting only the best estimator is a suboptimal strategy (using the N-best estimators typically yields higher phoneme recognition accuracies [8][18]), this test serves well as an indication of practical applications of our research. In addition, it also allows for positioning our techniques on the continuum between the best possible result (which exploits Oracle information) when the best stream is selected as the one, which gives the best result on known test data, and the accuracy obtained by selecting the stream for each sentence in random. Finally, we also report results obtained without any use of performance monitoring, i.e., the result obtained using all five bands of the full-band speech spectrum.

7. SUMMARY OF MAIN RESULTS

Because all measures are highly correlated with each other, we only show results of correlations with phoneme recognition error rate and choice of the best estimator for phoneme recognition. Furthermore, although we observed that the predictions improve significantly with the increasing length of test data used for prediction, we only show results using a single test sentence for prediction. Details of experiments mentioned here, comparisons with more investigated techniques and breakdown of results for different types of noises are reported elsewhere [17].

7.1. Evaluations by correlations with phoneme error rate of phoneme recognizer

400 correlation values are computed, one for each sentence of the test data, across the 31 phoneme recognizers using respective probability estimators. All five measures highly correlate with each other. The M-delta measure yields the highest correlations with phoneme accuracy in all conditions. The averaged correlation values for all data (clean and 9 different additive noises) from three evaluation techniques, the inverse entropy of the output of the estimator the previously proposed M-measure, and the new M-measure denoted here as delta M, are displayed in Figure 3 below.

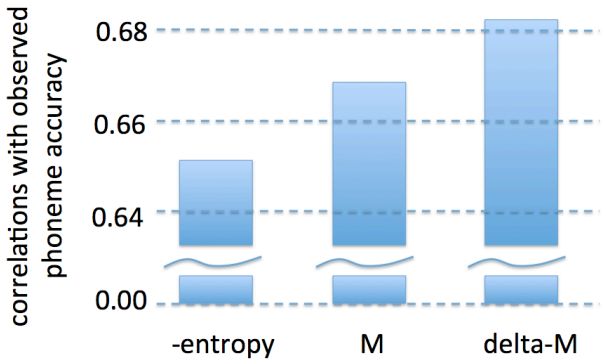

Fig. 3 Correlations with observed phoneme recognition accuracy

Techniques using ANN-based autoencoder or distributions of values on hidden layers of a ANN were also investigated. However, at this time, direct comparisons are not easily obtained, since the data, the estimators, and the noise conditions in these investigations were all different. However, these two techniques are topics of separate papers [11][12].

7.2. Evaluations by selecting processing streams in multistream speech recognition

Our second evaluation method uses the techniques developed for prediction of errors to select the most efficient phoneme recognizer out of the 31 streams. We compare the performance of the stream selected by the predictor the best possible result obtained using

Oracle knowledge (which in this case involves selecting the recognizer with the lowest phoneme error rate) and to a recognizer that uses the full speech band without any performance monitoring

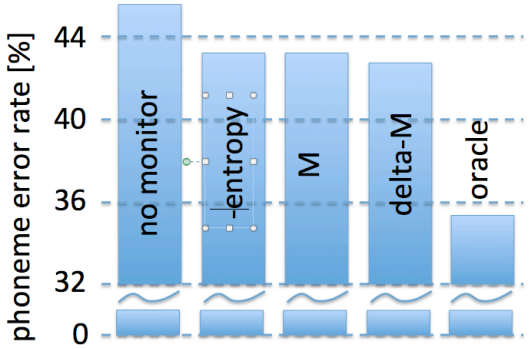

Fig. 4 Phoneme error rates of best estimators selected by different accuracy prediction techniques

Figure 4 summarizes some key results from the workshop. Using any of the performance monitors results in improvement over using the full 5 sub-band combination. Consistent with correlations reported above, the M-delta measure again yields the best results. Results of the oracle recognition indicates the space for improvement of performance monitoring techniques. For clarity, we show only recognition results using the best stream; some previous work [8][18] suggests that using the N-best streams considerably improves results.

7.3. Comparisons with conventional confidence estimation measures

We compare our results to results obtained using the measure based on inverse entropy of estimator output [5][6]. We also pursued the more conventional HMM-based raw acoustic score technique of confidence estimation (see, eg. [3]), where the averaged likelihood of the best path through the sequence of scaled likelihoods of subword models is used as a measure of the goodness of the subword posterior estimation. This technique yielded results that were similar to results obtained by inverse entropy technique.

So far, the newly introduced M-delta method yields consistent advantage compared to entropy-based technique when speech is corrupted by broad-band noise, and more significant gains when speech is corrupted by steady noise that is concentrated in narrow frequency range. This is advantage in narrow-band noise is understood since the narrow-band noise can yield high probability and low entropy estimates that are entirely wrong. However, such noise-induced estimates are revealed by the temporal-domain delta-M technique since they do not follow speech-like dynamics.

8. SOME REMAINING OPEN ISSUES

One issue that requires further investigation is the amount of speech data needed for reliable accuracy prediction. Our research so far mostly used estimation based on a single TIMIT sentence. Some of our preliminary evidence indicates that the accuracy of the prediction increases with up to tenths of sentences.

Measures of how well each predictor performs also require further attention. We strived for application-agnostic measures, and thus used correlations with known results. However, correlations are sensitive to nonlinear parameter transforms. Measures such as mutual information, which are invariant to parameter transforms, could be also applied.

Finally, similarity-based measures such as correlations evaluate the whole range of results and weigh all the good and the bad outputs equally in judging the success of a predictor. Care needs to be taken to better understand this issue and to provide the correct balance of the good and bad results to obtain a meaningful measure. In some other applications (e.g., selecting the N-best estimators) we need to know which estimators are bad. However, when the task is merely to get the best of the available estimators, we only need to predict the best outputs. Thus, ultimately, the evaluation needs to be directly linked to the intended application.

Selecting the best estimator among the pre-trained available ones is not the only way of adapting the estimator. Alternative adaptation techniques could be developed and applied with our techniques.

9. DISCUSSION AND SUMMARY

Most of current powerful machine learning techniques rely on large amounts of training data. Nevertheless, there are a number of practical situations where the recognizer encounters data from domains, which were not seen in training. Such a domain mismatch is significant problem in stochastic ASR.

Identifying the domain mismatch does not mean identifying data outliers. Instead, we need to identify data that that can cause problems for a particular probability estimator, and those two tasks are not always the same. Data could well be within the range of the previously seen training data but the information that they carry is for some reason corrupted, and subsequently the results of the estimation are also corrupted. In such situations it would be desirable for the data that cannot be well accommodated to be automatically identified as such and appropriately dealt with.

Our current efforts described in this paper represent progress towards this goal. The results presented here show that the unsupervised estimation of accuracy of a phoneme recognizer is possible. When successful, it could be applied in the adaptation loop of the recognizer and could provide a desirable alternative to the seemingly never-ending increases in amounts of training data.

We have proposed and investigated several new techniques for predicting accuracy of estimation of posterior probabilities of speech sounds on previously unseen data, and have shown the feasibility of this task. A number of the techniques were found to be effective in addressing the task. Due to space limitations, we could only briefly describe here some basic principles of the techniques that were investigated during the workshop, and summarize the performance of the most promising ones.

The overall best technique to date is the delta-M measure but more work is needed to fully evaluate the relative advantages and disadvantages of various techniques before any firm conclusions could be made. There is no doubt that new alternative techniques will emerge when the ASR community fully recognizes the importance of the problem.

REFERENCES

1. Geman, Stuart, Elie Bienenstock, and René Doursat. "Neural networks and the bias/variance dilemma." *Neural computation* 4.1 (1992): 1-58.

2. Kittler, Josef, et al. "Domain anomaly detection in machine perception: A system architecture and taxonomy." IEEE Trans. Pattern Analysis and Machine Perception, *in print.*

3. Seigel, Matthew Stephen, and Philip C. Woodland. "Combining Information Sources for Confidence Estimation with CRF Models." *Proceedings of* . *INTERSPEECH 2011*, pp. 905-908, International Speech Communication Association, 2011.

4. Tibrewala, Sangita, and Hynek Hermansky. "Sub-band based recognition of noisy speech." *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, IEEE,* 1997.

5. Okawa, Shigeki, Enrico Bocchieri, and Alexandros Potamianos. "Multi-band speech recognition in noisy environments." *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998*. Vol. 2. IEEE, 1998.

6. Misra, Hemant, Shajith Ikbal, Hervé Bourlard, and Hynek Hermansky. "Spectral entropy based feature for robust ASR." In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing,* vol. 1, pp. I-193*,* IEEE, 2004.

7. Mesgarani, Nima, Samuel Thomas, and Hynek Hermansky. "Adaptive Stream Fusion in Multistream Recognition of Speech." *Proc*eedings of *INTERSPEECH 2011*, International Speech Communication Association, 2011.

8. Variani, Ehsan, Feipeng Li, and Hynek Hermansky. "Multi-stream recognition of noisy speech with performance monitoring." *Proc*eedings of *INTERSPEECH 2013*, pp. 2978-2981. International Speech Communication Association, 2013.

9. Hermansky, Hynek, Ehsan Variani, and Vijayaditya Peddinti. "Mean temporal distance: Predicting ASR error from temporal properties of speech signal.", Proceedings of *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),*. IEEE, 2013.

10. Ogawa, Tetsuji, Harish Mallidi, Emmanuel Dupoux, Jordan Cohen, Naomi Feldman, Hynek Hermansky. "Delta-M measure for accuracy prediction and its application to multistream-based unsupervised adaptation", *submitted*.

11. Cohen, Jordan R. "Segmenting speech using dynamic programming." *The Journal of the Acoustical Society of America* 69.5, pp. 1430-1438, 1981.

12. Mallidi, Harish, Tetsuji Ogawa, Karel Vesely and Hynek Hermansky. "Autoencoder based performance monitoring", *submitted.*

13. Veselý, Karel, Harish Mallidi, Vijayaditya Peddinti, Tetsuji Ogawa, Lukas Burget and Hynek Hermansky. "Towards hidden performance predictor", *submitted.*

14. Tomar, Vikrant Singh, and Richard C. Rose. "Manifold Regularized Deep Neural Networks. " *Proceedings of INTERSPEECH* 2014, pp. 348-353, International Speech Communication Association, 2014.

15. Hermansky, Hynek, Sangita Tibrewala, and Misha Pavel. "Towards ASR on partially corrupted speech." *Proceedings of Fourth International Conference on Spoken Language Processing ICSLP 96.* Vol. 1. IEEE, 1996.

16. Bourlard, Hervé, and Stéphane Dupont. "A new ASR approach based on independent processing and recombination of partial frequency bands.". *Proceedings of Fourth International Conference on Spoken Language Processing ICSLP 96,* Vol. 1. IEEE, 1996.

17. Hermansky, Hynek. "Multistream Recognition of Speech: Dealing With Unknown Unknowns." *Proceedings of the IEEE* 101.5 (2013): 1076-1088.

18. Ogawa, Tetsuji, Feipeng Li, and Hynek Hermansky. "Stream selection and integration in multistream ASR using GMM-based performance monitoring," in *Proc. INTERSPEECH 2013*, pp. 3332-3336, International Speech Communication Association, 2013.