TRAPPING CONVERSATIONAL SPEECH: EXTENDING TRAP/TANDEM APPROACHES TO CONVERSATIONAL TELEPHONE SPEECH RECOGNITION

Nelson Morgan[1,2] *Barry Y. Chen*[1,2] *Qifeng Zhu*[1] *Andreas Stolcke*[1,3]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] University of California Berkeley, Berkeley, CA, USA
[3] SRI International, Menlo Park, CA, USA
{morgan, byc, qifeng, stolcke}@icsi.berkeley.edu

ABSTRACT

TempoRAl Patterns (TRAPs) and Tandem MLP/HMM approaches incorporate feature streams computed from longer time intervals than the conventional short-time analysis. These methods have been used for challenging small- and medium-vocabulary recognition tasks, such as Aurora and SPINE. Conversational telephone speech recognition is a difficult large-vocabulary task, with current systems giving incorrect output for 20-40% of the words, depending on the system complexity and test set. Training and test times for this problem also tend to be relatively long, making rapid development quite difficult. In this paper we report experiments with a reduced conversational speech task that led to the adoption of a number of engineering decisions for the design of an acoustic front end. We then describe our results with this front end on a full-vocabulary conversational telephone speech task. In both cases the front end yielded significant improvements over the baseline.

1. AUGMENTING CONVENTIONAL FEATURES

For decades, the feature extraction component of speech recognition engines has consisted of some form of local spectral envelope estimation, typically with some simple transformation; current front ends are based largely on the Mel cepstrum or perceptual linear prediction (PLP) [1] computed from an analysis window of roughly 25 or 30 ms surrounding a central signal point, stepped along every 10 ms. A number of alternatives have been developed in recent years. One such approach, tandem acoustic modeling [2, 3, 4] uses a multi-layer perceptron (MLP) to first discriminatively transform multiple feature vectors (typically PLP from 9 frames) before using them as observations for Gaussian mixtures hidden Markov models (GMHMM). Thus, the neural network, which could be called a "feature net", incorporates around 100 ms of speech. In this paper we will refer to the resulting variables as PLP/MLP features. Others have also tried incorporating longer temporal information yielding significant improvements in speech recognition performance (e.g., [5]).

The MLP is typically trained using phonetic targets. This approach works very well in matched training and test conditions, often achieving lower word error rates than systems without the discriminant nonlinear transformation provided by the MLP. However, in the case of mismatched training and testing conditions, ICSI and OGI researchers working on the Aurora task found it preferable to augment the original features with the feature net outputs, essentially using the concatenation of the original features and the PLP/MLP features as the front end for the GMHMM [6]. A similar approach was used in [7], where standard features were augmented by a complimentary source of information (in this case, estimates of formants from a mixture of Gaussians).

Another promising approach has been to combine the PLP/MLP features with features derived from the outputs of MLPs incorporating long-time log critical band energy trajectories (500 ms - 1 s) [8, 9]. The set of these MLPs forms the TRAPS system, named as such because the system learns discriminative TempoRAl Patterns (TRAPS) in speech. MLPs in the TRAPS system are also trained with phonetic targets. We have observed that systems using the combination of the two feature sets perform better than those using either feature type alone.

The approaches listed above were developed on small tasks, i.e., connected digits, continuous numbers, and TIMIT phone recognition, where the training and testing sets were small in both vocabulary and data size. We have now tested systems that incorporate these features in two progressively larger tasks. We used conventional front end features (12th order PLP plus energy and derivatives), augmented with the combination of PLP/MLP and TRAPS features. These corresponded to three different temporal spans. The original PLP features were derived from short term spectral analysis (25 ms time slices every 10 ms). In contrast, PLP/MLP used 9 frames of PLP features (100ms), and TRAPS used 51 frames of log critical band energies (500ms). For the PLP/MLP stream, we trained discriminative feature net MLPs using 46 phoneme targets generated from forced alignments using the SRI DECIPHER recognizer. For the second stream, the first stage TRAPS MLPs took log critical band energy trajectories, formed by taking 51 consecutive frames of log critical band energies every 10ms, and transformed by principal component analysis (PCA). These critical band MLPs were trained with the same phoneme targets as in the feature net MLP. A "merger" MLP (trained with these same phoneme targets) combined the output of the critical band MLPs to produce a single estimate of phoneme posteriors every 10 ms.

Since the outputs of both the TRAPS classifier and the PLP net can be interpreted as posterior probabilities of the 46 phonemes, we could combine them using frame-wise posterior probability combination techniques [10, 11] (described briefly below). After combination, we took the log of the posterior vector to make it more Gaussian, and then orthogonalized and reduced the dimensionality of the posterior vector using PCA. The resulting variables were then appended to the original PLP cepstra to form the augmented feature vector. Figure 1 summarizes this process.

In what follows, we refer to these augmented feature vectors

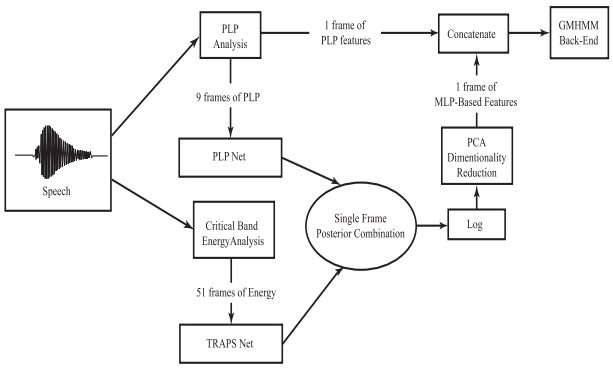

Fig. 1. *Augmenting PLP Front End Features*

as PLP+*combomethod*(*Streams*), where *combomethod* can be one of three frame-wise posterior combination methods: the average of the posteriors combination (AVG); the average of log posteriors combination (AVGLog), and finally, the inverse entropy weighted combination (INVENT) [11]. *Streams* refers to the PLP/MLP feature stream and TRAPS feature stream. The first two combination methods essentially assume that each MLP feature stream is equally important, while the entropy-based combination assumes that the MLP feature with lower entropy is more important than an MLP feature with high entropy. This is intuitively correct, since a low entropy posterior distribution (such as would occur with a high single peak) implies strong confidence in class identity. Generally, the combined posterior can be written as:

$$\tilde{P}(q_k|x) = \omega_1 \tilde{P}(q_k|x^1) + \omega_2 \tilde{P}(q_k|x^2) \quad (1)$$

where $\tilde{P}(q_k|x^1)$ and $\tilde{P}(q_k|x^2)$ are the posteriors (or log posterior in the case of log posterior average combination) from two different MLPs for the same frame k. In both variants of average combination, ω_1=ω_2=0.5. For entropy-based posterior combination, ω is the inverse entropy computed over one frame for an MLP output and normalized so that the sum of all weights is one. We also use a threshold in our entropy computations as done in [11], so that if the entropy for a frame from an MLP is greater than 1, it is set to a large value (e.g., 10000).

In all of our initial experiments we performed, the baseline feature vector consisted of 12th order PLP coefficients plus energy, along with 1st and 2nd order deltas, to yield a 39-dimensional baseline feature. We also used mean and variance normalization per conversation side.

We used a stripped-down version of SRI's Hub-5 conversational speech transcription system for our HMM back-end. In particular, the back-end that we used was similar to the first pass of the system described in [12], using a bigram language model and within-word triphone acoustic models.

2. THE 500 WORD CTS TASK

Before applying our approaches to the full-vocabulary Switchboard task, we considered a more limited task, that of recognizing the 500 most common words[1] in Switchboard I, the most commonly used conversational telephone speech (CTS) corpus. Given the frequent occurrence of these words, it was likely that error rate reduction would also apply to Switchboard in general, and less training data would be required than would be needed for the full task. This in turn sped training time accordingly. Finally, decoding complexity for this task was smaller, which also improved experiment turn-around times.

Table 1. Training set composition details

Data Source	Training set size (hours of data)			
	"Short"		RUSH	
	Male	Female	Male	Female
English CallHome	0.56	2.75	0.19	0.92
Mississippi SWB1	30.28	31.30	10.08	10.63
SWB Cellular	1.83	2.03	0.59	0.69
SWB Credit Card	0.20	0	0.06	0
Total	32.87	36.08	10.92	12.24

For training our 500-word system, we created a subset of the "Short" training set used at SRI for CTS system development, which we referred to as the Random Utterances of Short Hub or the RUSH set. The "Short" training set takes telephone speech data from four sources: English CallHome, Switchboard I with transcriptions from Mississippi State [13], Switchboard Cellular, and Switchboard Credit Card Corpus. Our RUSH subset randomly picks one third of the total number of utterances spoken by each speaker in the "Short" training set. Table 1 describes the composition of both "Short" and RUSH training sets.

The 500-word test set was a subset of the 2001 Hub-5 evaluation data. Given the 500 most common words in Switchboard I, we chose utterances[2] from the 2001 evaluation data in which at most 10% of the words in an utterance were out-of-vocabulary (OOV) words. 49.6% of the utterances in the 2001 evaluation data met this requirement, and the total OOV rate on the retained utterances was 3.2%. We then partitioned this set into a tuning set (0.97 hours, 8,242 word tokens) and a test set (1.42 hours, 11,845 word tokens). We used the tuning set to adjust the word transition weight and language model scaling, and we determined word error rates on the test set. The language model (LM) used was the first-pass bigram used by SRI for Hub-5 evaluations in 2000. Note that, although the vocabulary of the test set was limited by virtue of the data selection process, the recognition LM was the same 33k-word vocabulary used in the standard recognition system, so that OOV words in the test set were not a significant source of error.

2.1. Results on Top 500 Words Task

Using the baseline PLP features, we trained gender-dependent triphone HMMs on the 23-hour RUSH training set, and then tested this system on the 500-word test set, achieving a 43.8% word error rate (see Table 2). As seen in the table, the word error rate was reduced 10% relative by augmenting the baseline features with the gender-dependent PLP/MLP and TRAPS features.

The three combination methods yielded similar word error rates, though the inverse entropy approach was slightly better (but not by a statistically significant margin). The averaging methods have the advantage of simplicity, and don't rely on any estimation method. On the other hand, we observed that the inverse entropy combination technique was sometimes robust to poor classifier

[1] This reduced task was proposed by our colleague George Doddington.

[2] We treated each of the waveform segments defined by NIST for evaluation purposes as one utterance, regardless of whether these represented coherent linguistic units or not.

Table 2. Word error rate (WER) and relative reduction of WER on the top 500-word test set of systems trained on the RUSH set using different combination approaches. *Streams* denotes the PLP/MLP feature stream and the TRAPS feature stream.

Feature Vector	500 Word Test Set WER	Relative Reduction WER
PLP Baseline	43.8%	-
PLP+AVG(*Streams*)	39.4%	10.0%
PLP+AVGLog(*Streams*)	39.5%	9.8%
PLP+INVENT(*Streams*)	39.2%	10.5%

streams. In one experiment, we unintentionally combined a badly degraded TRAPS stream with the other features using both methods. When probabilities were multiplied or added without weights, the degraded stream hurt performance badly. On the other hand, the inverse entropy-weighting automatically reduced the importance of the poor stream so that the overall performance essentially matched what was achieved for a feature vector that consisted of the baseline PLP features concatenated with the PLP/MLP feature alone. Thus, the entropy-based approach to combination appears to be more robust to unexpectedly poor streams. This property might be particularly useful for future efforts in which we might combine a larger number of streams, most of which will provide less useful information for any particular frame.

3. FULL CTS VOCABULARY

Since the new front end improved performance on the 500-word task, we incorporated it in a system that was tested on the full-vocabulary CTS task.

Error rates on Switchboard test sets were unacceptably high for systems trained on the RUSH training set alone, so we used SRI's "Short" CTS training set from which RUSH was derived. See Table 1 for the "Short" training set composition. As in the 500-word task, we trained triphone gender-dependent HMMs as well as gender-dependent PLP/MLP feature nets and TRAPS systems.

For testing, we used the full 2001 Hub-5 Switchboard evaluation set. This evaluation set contains a total of 6.33 hours of speech, with 62,890 total word tokens. For tuning our system parameters, we used a subset of the disjoint 2001 Hub-5 development set.

3.1. Results on Full CTS Task

The baseline system achieved a 43.8% word error rate on the 2001 Hub-5 evaluation set (see Table 3). The augmented features reduced the error rate by about 7% relative. The last row in the table shows an improved result for the INVENT stream combination approach, after the system was further tuned using an SRI DECIPHER parameter called the "Gaussian weight", which in the earlier experiments had been optimally tuned for the baseline feature. The Gaussian weight scales the Gaussian log likelihoods but does not affect other parts of the acoustic model (mixture weights, transition probabilities). With more feature dimensions, the weight is typically tuned to a smaller number, thus compensating for the added terms in the Gaussian likelihoods and the fact that feature dimensions are not independent despite the use of diagonal covariances.

Table 3. Word error rate (WER) and relative reduction of WER on the 2001 Hub-5 evaluation set of systems trained on SRI's "Short" CTS training set using different combination approaches. *Streams* denotes the PLP/MLP feature stream and the TRAPS feature stream. The last row corresponds to a version that has been optimized for the best Gaussian weight for the new feature vector; the baseline had already been tuned in this way.

Feature Vector	Hub-5 EVAL2001 WER	Relative Reduction WER
PLP Baseline	43.8%	-
PLP+AVG(*Streams*)	40.5%	7.5%
PLP+AVGLog(*Streams*)	41.0%	6.4%
PLP+INVENT(*Streams*)	40.6%	7.3%
PLP+INVENT(*Streams*)'	39.6%	9.6%

With this additional tuning, the error rate reduction increased to almost 10% relative.

Finally, for these experiments, there was a small penalty for the AVGLog combination method in comparison to the other approaches.

4. PERFORMANCE USING A BETTER BASELINE

While the previous experiments are encouraging, it is important to see if the new front end will still yield comparable relative reductions in word error rates in the presence of other improvements that are typical for the current best research systems. In particular, since the new front end is discriminatively trained, a key question is whether other discriminant approaches such as heteroskedastic linear discriminant analysis (HLDA) [14], which is now commonly used in such systems, might not provide error reductions that would be redundant with those provided by the proposed front end. More generally, we wanted to significantly improve the baseline system by other means in order to see if the error rate reduction from the proposed front end was still significant.

To explore these issues, we repeated the task described in Section 3, with the same training and testing sets. The baseline feature used in this section is the PLP and energy feature with the first three derivatives, transformed by HLDA. HLDA is a discriminant transformation that is used to reduce the feature dimension from, in our case, 52 to 39. As implemented in the SRI system, the transform is trained to optimize the discrimination among the Gaussians in a reference model of phonetically-tied mixtures previously trained on the same data. It has been observed by several research groups that adding the third cepstral derivatives can be useful when followed by HLDA. We will refer to this new baseline feature as HLDA(PLP_ddd). The new baseline system also uses a significantly improved bigram LM, incorporating more sources, improved smoothing, and a total of 3.2 million bigrams as opposed to 1.3 million bigrams in the earlier system. While there is no obvious redundancy between an improved LM and an improved front end, it is nonetheless likely that some of the same errors are prevented by each of these improvements.

We augmented HLDA(PLP_ddd) with an MLP-based feature vector similar to the one described in Section 1. Using the same MLP/PLP and TRAPS streams as before, we combined their posterior outputs using inverse entropy weighting. We performed

Table 4. Word error rates (WER) and relative reduction of the WER on the 2001 Hub-5 evaluation set with the improved baseline and augmented features.

Feature Vector	Hub5 EVAL2001 WER	Relative Reduction WER
HLDA(PLP_ddd)	37.2%	-
HLDA(PLP_ddd) +HLDA(INVENT(streams))	34.4%	7.5%

mean and variance normalization by conversation side on the MLP-based features, as was done for the baseline features. Finally, HLDA (instead of previously PCA) was applied to the MLP feature, reducing it to 25 dimensions. The same HLDA training criterion and procedure as for the HLDA(PLP_ddd) feature was employed We call the resulting feature HLDA(INVENT(streams)), and concatenated it with HLDA(PLP_ddd). As for the last result in Table 3, the Gaussian weight parameter was tuned (on the independent tuning set).

The word error rates of systems using baseline and augmented feature vectors are shown in Table 4. As noted in the table, the new features provided significant reductions in error (7.5% relative). The error reduction relative to the improved baseline with HLDA and improved LM is slightly smaller, but on the same order as that obtained with the previous baseline.

5. DISCUSSION AND CONCLUSIONS

As we had hoped, incorporating the augmented feature vector significantly reduced the word error rate for both reduced and full-vocabulary tasks. In both cases the combination methods all seemed to be effective. However, the inverse entropy method was the best (or nearly so) for both tasks, and also appeared to be robust to catastrophic degradations of feature streams. This property may be more important if we apply these methods to combining a larger number of probabilistic feature streams. This view seems to be supported by earlier work at IDIAP [11]. Using dynamically weighted combination of streams at the feature level may also be advantageous for features that are only occasionally useful. Such features will tend to be be of little help when combined at higher levels (e.g., for word candidates with ROVER).

The PLP/MLP and TRAPS features that were used here were significantly different from the baseline features. As one might expect, this complementarity leads to improvements. Furthermore, the improvements in the smaller task were roughly predictive of improvements in the larger task. On the other hand, it is likely that further optimization of performance can be achieved by work with the larger task. This is always true, but the ability to bring some of the performance improvements forward following work on a smaller task is extremely important for speeding the development of novel approaches. Our experience suggests that providing an intermediate task as a "stepping stone" can greatly aid in development of new speech recognition modules, particularly for the front end. Our next step is to incorporate these methods in one or more of the current best laboratory systems for conversational telephone speech.

6. ACKNOWLEDGMENTS

We would like to gratefully acknowledge all the people who helped provide various components for our system: Sunil Sivadas and Hynek Hermansky for their help in providing the TRAPS features for the 500-word and full-vocabulary tasks; Hemant Misra and Hervé Bourlard for providing the inverse entropy technique; and George Doddington for finding the top 500 words in Switchboard and creating the top 500-word subset from the 2001 Hub-5 evaluation set. All of the other members of our EARS Novel Approaches team (at SRI, OGI, IDIAP, Columbia, and the University of Washington) also contributed intellectually to this work. Finally, this work was made possible by funding from the DARPA EARS Novel Approaches Grant: No. MDA972-02-1-0024.

7. REFERENCES

[1] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, April 1990.

[2] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. ICASSP-2000*, Istanbul, vol. III, pp. 1635–1638.

[3] D.P.W. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proc. ICASSP-2001*, Salt Lake City, May.

[4] D.P.W. Ellis and M.J. Reyes Gomez, "Investigations into tandem acoustic modeling for the Aurora task," in *Proc. Eurospeech-2001, Special Event on Noise Robust Recognition*, Denmark, September.

[5] B. Milner, "Inclusion of temporal information into features for speech recognition," in *Proc. ICSLP-1996*, pp. 256–259.

[6] C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, and S. Sivadas, "Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks," in *Proc. Eurospeech-2001*, Denmark, September.

[7] M. N. Stuttle and M. J. F. Gales, "A mixture of Gaussians front end for speech recognition," in *Proc. Eurospeech-2001*, Denmark, September.

[8] H. Hermansky and P. Sharma, S.and Jain, "Data-derived nonlinear mapping for feature extraction in HMM," in *Proc. ICASSP-2000*, Istanbul.

[9] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. ICASSP-1999*, Phoenix, March.

[10] A. Janin, D. Ellis, and N. Morgan, "Multi-stream speech recognition: Ready for prime time?," in *Proc Eurospeech-1999*, Budapest, vol. II, pp. 591–594.

[11] H. Misra, H. Bourlard, and V. Tyagi, "New entropy based combination rules in HMM/ANN multi-stream ASR," in *Proc. ICASSP-2003*, Hong Kong.

[12] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.

[13] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," in *Proc. ICSLP-1998*, Sydney, Australia, November, pp. 1543–1546.

[14] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions Speech and Audio Pro-cessing*, vol. 7, pp. 272–281, 1999.