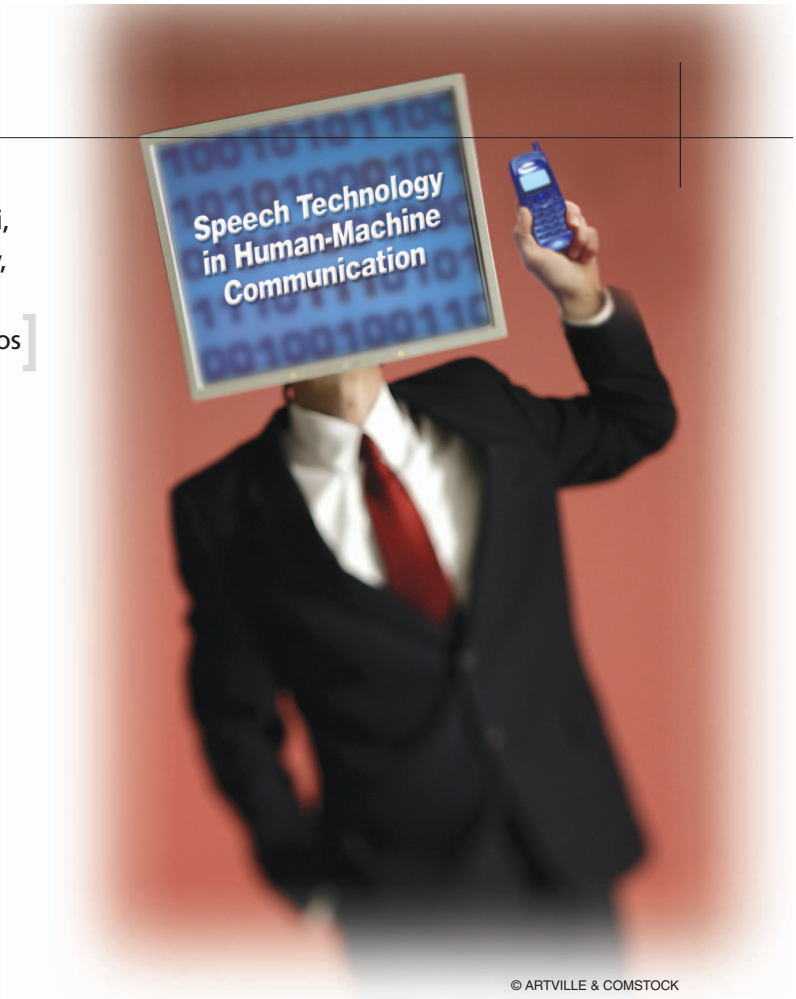
Nelson Morgan, Qifeng Zhu, Andreas Stolcke, Kemal Sönmez, Sunil Sivadas, Takahiro Shinozaki, Mari Ostendorf, Pratibha Jain, Hynek Hermansky, Dan Ellis, George Doddington, Barry Chen, Özgür Çetin, Hervé Bourlard, and Marios Athineos

Beyond the spectral envelope as the fundamental representation for speech recognition

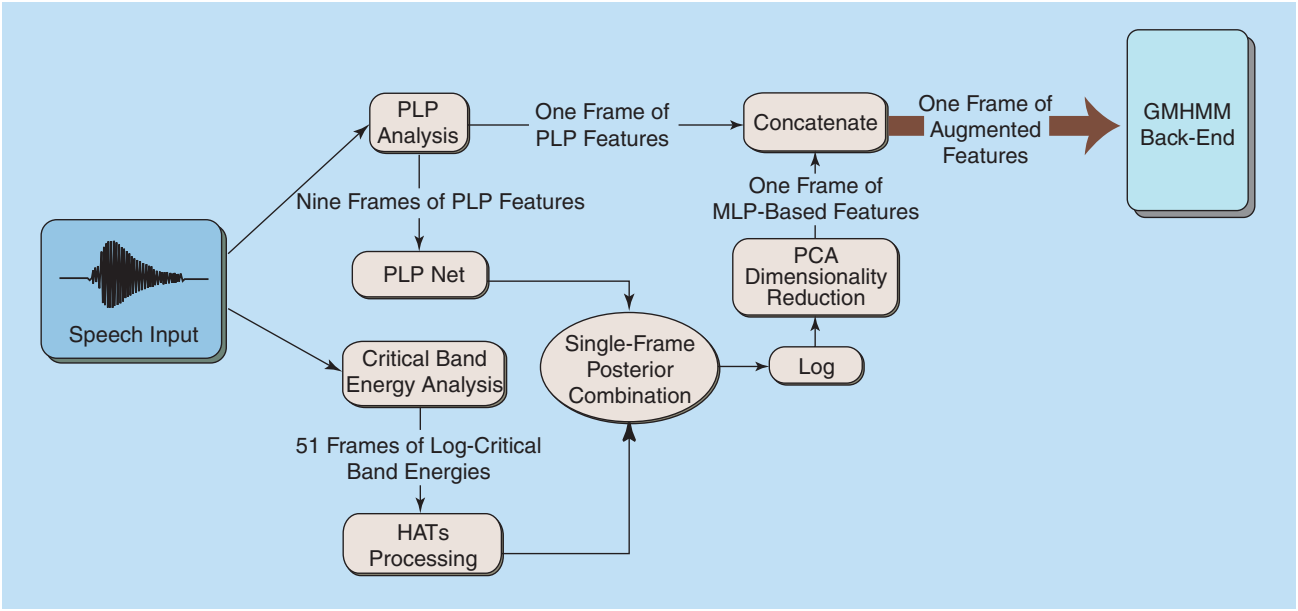


Pushing the Envelope—Aside

State-of-the-art automatic speech recognition (ASR) systems continue to improve, yet there remain many tasks for which the technology is inadequate. The core acoustic operation has essentially remained the same for decades: a single *feature vector* (derived from the power spectral envelope over a 20–30 ms window, stepped forward by ~10 ms per frame) is compared to a set of *distributions* derived from training data for an inventory of subword units (usually some variant of phones). While many systems also incorporate time derivatives [16] and/or projections from five or more frames to a lower dimension [21], [17], the fundamental character of the acoustic features has remained quite similar. We believe that this limited perspective is a key weakness in speech recognizers. Under good conditions, human phone error rate for nonsense syllables has been estimated to be as low as 1.5% [1], as compared with rates that are over an order of magnitude higher for the best machine phone recognizers [13], [23], [26]. In this light, our best current recognizers appear half deaf, only making up for this deficiency by incorporating strong domain constraints. To develop generally applicable and useful recognition techniques, we must overcome the limitations of current acoustic processing. Interestingly, even human phonetic categorization is poor for extremely short segments (e.g., <100 ms), suggesting that analysis of longer time regions is somehow essential to the task. This is supported by information theoretic analysis, which shows discriminative dependence conditional on underlying phones between features separated in time by up to several hundred milliseconds [6], [28].

In mid-2002, we began working on a Defense Advanced Research Projects Agency (DARPA) sponsored project known as the "Novel Approaches" component of the Effective Affordable Reusable Speech-to-text (EARS) program. The fundamental goal of this multisite effort was to "push" the spectral envelope away from its role as the sole source of acoustic information incorporated by the statistical models of modern speech

1053-5888/05/$20.00©2005IEEE



[FIG1] Posterior-based feature generation system. Each posterior stream is created by feeding a trained multilayer perceptron (MLP) with features that have different temporal and spectral extent. The "PLP Net" is trained to generate phone posterior estimates given roughly 100 ms of telephone bandwidth speech after being processed by PLP analysis over nine frames. HATs processing is trained for the same goal given 500 ms of log-critical band energies. The two streams of posteriors are combined (in a weighted sum where each weight is a scaled version of local stream entropy) and transformed as shown to augment the more traditional PLP features. The augmented feature vector is used as an observation by the Gaussian mixture hidden Markov model (GMHMM) system.

recognition systems (SRSs), particularly in the context of the conversational telephone speech recognition task. This ultimately would require both a revamping of acoustical feature extraction and a fresh look at the incorporation of these features into statistical models representing speech. So far, much of our effort has gone towards the design of new features and experimentation with their incorporation in a modern speech-to-text system. The new features have already provided significant improvements in such a system in the 2004 NIST evaluation of recognizers of conversational telephone speech. The development of statistical models to best incorporate the long time features is being explored, but development is still in its early stages.

BACKGROUND

Mainstream speech recognition systems typically use a signal representation derived from a cepstral transformation of a short-term spectral envelope. This dependence on the spectral envelope for speech sound discrimination dates back to the 1950s, as described in [11]. In turn, this style of analysis can be traced back to the 1930s vocoder experiments of Homer Dudley [14]. Perhaps more fundamentally, many speech scientists have observed the relationship between the spectral components of speech sounds and their phonetic identity. They have further characterized these sounds by their correspondence to the state of the speech articulators and the resulting resonances (formants). By this view, one should use pattern recognition techniques to classify new instances of speech sounds based on their proximity in some spectral (or cepstral) space to speech sounds collected for training the system. Modern statistical speech recognition systems are fundamentally elaborations on this principle; individual training examples are not used directly for calculating distances but rather are used to train models that represent statistical distributions. The Markov chains that are at the heart of these models represent the temporal aspect of speech sounds and can accommodate differing durations for particular instances. The overall structure provides a consistent mathematical framework that can incorporate powerful learning methods such as maximum likelihood training using expectation maximization [12]. Systems using short-term cepstra for acoustic features and first-order Markov chains for the acoustic modeling have been successful both in the laboratory and in numerous applications, ranging from cell phone voice dialing to dialog systems for use in call centers.

Despite these successes, there are still significant limitations to speech recognition performance, particularly for conversational speech and/or for speech with significant acoustic degradations from noise or reverberation. For this reason, we have proposed methods that incorporate different (and larger) analysis windows, which will be described below. We note in passing that we and many others have already taken advantage of processing techniques that incorporate information over long time ranges, for instance for normalization (by cepstral mean subtraction [2] or relative spectral analysis (RASTA) [18]). We also have proposed features that are based on speech sound class posterior probabilities, which have good properties for both classification and stream combination.

TEMPORAL REPRESENTATIONS FOR EARS

Our goal is to replace (or augment) the current notion of a spectral-energy-based vector at time t with variables based on

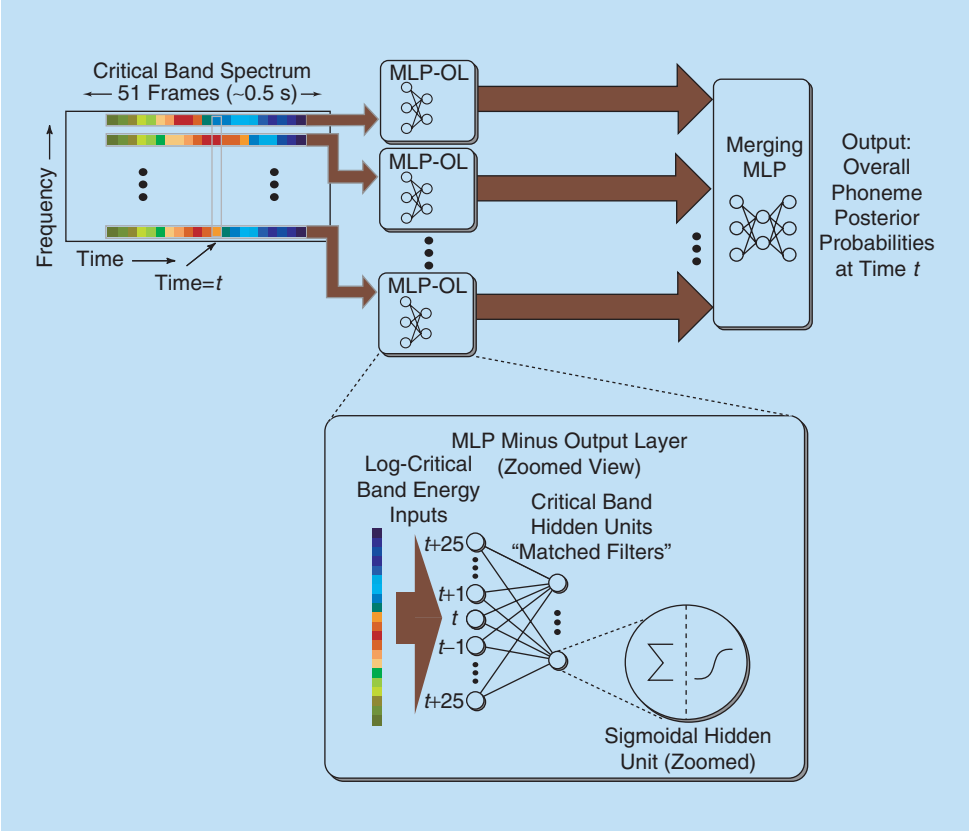
posterior probabilities of speech categories for long and short time functions of the time-frequency plane. These features may be represented as multiple streams of probabilistic information. Other analyses that use time and frequency windows will intermediate between these extremes. For all the reasons described earlier, it is desirable to extend the acoustic analysis to significantly larger time windows than are typically used.

As a specific example, we are working with narrow spectral subbands and long temporal windows (up to 500 ms or more, sufficiently long for two or more syllables). In one successful instantiation of these ideas, we used the temporal trajectory of logarithmic spectral energy in a single critical band as the input vector for a multilayer perceptron, which then generates class posterior probability estimates. This structure is referred to as TempoRAl patterns (TRAPs) [19]. We developed this approach for use in a multistream combination system alongside conventional features, using simple functions of the posterior estimates as intermediate features. This yielded promising results for small tasks like numbers recognition, as reported in [20]. In recent years, we have begun to apply related techniques to the recognition of conversational telephone speech. As reported in [10], a variant of this approach called hidden activation TRAPs (HATs) has led to relative reductions in error of around 8% when these features are combined with posterior estimates from perceptual linear prediction (PLP) and its first two derivatives and incorporated in the SRI decipher system (reducing the word error rate (WER) from 25.6% to 23.5% on the 2001 NIST evaluation set). These results carried over to the full SRI evaluation system as well. In fact, on the 2004 NIST evaluation set, the system incorporating these features achieved an 18.3% WER, a 9.9% reduction in error from the 20.3% achieved by the same system without these new features.

The subsystem for feature generation that achieved these results is shown in Figure 1, and the HATs component is illustrated in Figure 2.

In addition to the large-vocabulary recognition work mentioned earlier, we have also used small vocabulary tasks to further develop the temporal features beyond those based on the sequence of framewise log energies used by the TRAPs and HATs approaches. We have designed features based on linear prediction coefficients calculated on the spectrum, which form a mathematical dual to the more familiar time-domain linear predictive coding (LPC) models; we call this frequency-domain linear prediction (FDLP) [3]. These features constitute a parametric model of the temporal envelope without any frame-rate subsampling. They can capture fine temporal detail about transients that is lost in conventional frame-based features. While the poles in conventional LPC can describe sharp spectral resonances very precisely, in the dual domain the poles of the linear predictor can be taken as parametric descriptions of individual sharp peaks in the temporal (Hilbert) envelope, without any implicit envelope smoothing or downsampling. Such transients occur during stop bursts and other rapid articulation

TO DEVELOP GENERALLY APPLICABLE AND USEFUL RECOGNITION TECHNIQUES, WE MUST OVERCOME THE LIMITATIONS OF CURRENT ACOUSTIC PROCESSING.



[FIG2] Hidden activation TRAP component. The posterior stream is created by feeding a "merging" multilayer perceptron (MLP) with the output from the hidden activation layers of three-layer MLPs that were each trained on a different critical band spectrum (referred to as "MLP-OL" or MLP minus the output layer in the figure).

changes in speech. This parametric representation (which has the useful property of automatically selecting the most "important" transients within its analysis window when fitting a limited-order model) allows us to investigate the importance of this aspect of the signal in speech recognizers for the first time.

We have experimented with different methods of incorporating this information into a conventional recognizer. Our best results come from modeling the envelope in four octave-wide frequency bands over relatively long windows (300 ms) and then selecting the poles that represent temporal features in the center of the window and using their individual magnitudes as values. Augmenting our baseline PLP system with these features reduced the WER on a natural numbers recognition task (Numbers-95) from 5.0% to 3.8% (a 24% relative improvement). In subsequent work, we incorporated the FDLP model into the TRAP framework, yielding the LP-TRAP [4]. Here we modeled the temporal envelope of 15 critical bands over longer windows (1 s). The all-pole polynomials were converted to cepstral coefficients for recognition. Using this method, we improved the WER of the spectrogram-based TRAP baseline from 5.9% to 5.3% (a 10% relative improvement) on the Numbers-95 task. Although both error rates were worse than the PLP-based system (since PLP has a number of advantages over purely spectral-based front ends), this initial result suggested that the all-pole polynomials might be a good way to represent spectral trajectories as part of a larger system.

MAINSTREAM SPEECH RECOGNITION SYSTEMS TYPICALLY USE A SIGNAL REPRESENTATION DERIVED FROM A CEPSTRAL TRANSFORMATION OF A SHORT-TERM SPECTRAL ENVELOPE.

STABILITY OF RESULTS

It is standard fare for conference papers to report the success of method Y (in comparison to method X) on some particular test. In the best cases, this result follows from some principles that are likely to apply to future cases. However, it is particularly gratifying if the observed results hold over a range of cases and if one can observe phenomena (other than final score on a single task) that appear to demonstrate some reason for the utility of the new method. We believe that we can now make such claims for the incorporation of long-term, posterior-based features ASR. To begin with, we have observed improvements for a range of conversational telephone speech recognition systems, with increasing amounts of training data and corresponding system complexity (and reduced baseline error rate) as indicated in Table 1. "Switchboard" and "Fisher" refer to earlier and later data collection approaches for the data that was distributed by the Linguistic Data Consortium (LDC). Note that the Switchboard and Fisher conversational data is extremely difficult to recognize due to their unconstrained vocabulary, speaking style, and range of telephones used. In addition to the consistent WER reductions provided by the new features, we also observed improved low-level classification performance for long-duration speech sounds like diphthongs. All of this work actually began with experiments on very different (and much smaller) tasks, in particular on recognition of natural numbers (e.g., 8, 18, 80, etc.). Results for these experiments were also consistent with what we observed for the more difficult conversational speech recognition task. As noted above for the newer FDLP and LP-TRAP approaches, we continue to do early experiments with small tasks to permit many experiments, but we take the best (and most developed) of these methods and validate their generality on large tasks.

[TABLE 1] RESULTS FROM A SERIES OF EXPERIMENTS WITH INCREASINGLY COMPLEX BASELINE SYSTEMS USING INCREASINGLY LARGE AMOUNTS OF TRAINING DATA.

TRAINING DATA	TEST SET	BASELINE WER %	RELATIVE REDUCTION IN WER % ADDING LONG-TERM FEATURES
SWITCHBOARD, 23 H	1.4 H SUBSET OF 2001 NIST EVAL SET	43.8%	10.5%
SWITCHBOARD, 64 H	2001 NIST EVAL SET	39.1%	10.0%
SWITCHBOARD + "FISHER," 200 H MALE ONLY	2001 NIST EVAL SET	30.8%	7.1%
SWITCHBOARD + FISHER, 2,000 H	2004 NIST EVAL SET	20.3%	9.9%

The first results row corresponds to experiments from a reduced task using utterances primarily composed of the most frequent words [9]. Both the Switchboard and Fisher collections consisted of impromptu conversations between randomly selected volunteers that were, however, sparked by suggested topics. A key practical difference between the data sets is that the Switchboard data was carefully transcribed, while the later Fisher set was transcribed with a quicker approach that was not quite as accurate.

SOME PRACTICAL CONSIDERATIONS

At this point, we have consistently seen the advantage in augmenting cepstral features with features that span much longer stretches of time, using discriminatively trained neural networks to transform the raw representations into posterior probabilities of phonetic classes and constraining the network trainings to emphasize the temporal trajectories of narrowband components of the speech signal. A number of difficulties remain, however. As we have experimented with larger and larger training sets, we have found that our new

features provide the best improvement when they are trained using the same amount of data that has been used to train the hidden Markov models (HMMs). However, we also achieved our best improvements when the network size grew proportionately with the training data. This implies a quadratic growth in training time, which can be a great burden for experimentation. For our current experiments, we have addressed this problem through a combination of small remedies: the use of hyperthreading on dual CPUs, gender-specific training, preliminary network training passes with fewer training patterns, and customization of the learning regimen to reduce the number of epochs. Despite the success of these measures, in the long term it may become important to find ways to improve the scaling properties of these methods, for instance by using selected subsets of the data for later training passes. Extending to clusters with more CPUs can help to some extent, though communication speed between processors limits the effectiveness of this approach. For the moment, however, our methods have proven sufficient to scale up from tens of hours to thousands of hours of training material. Figure 3 illustrates the key idea behind these methods, namely, the incorporation of smaller amounts of training material for early, large-learning-rate passes.

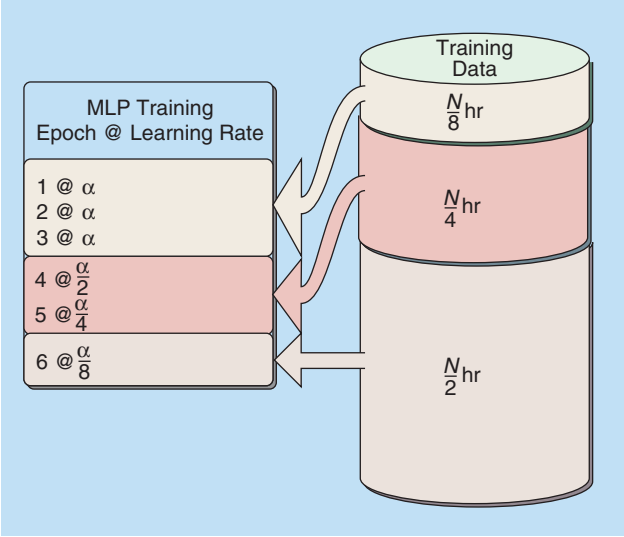
AT THIS POINT, WE HAVE CONSISTENTLY SEEN THE ADVANTAGE IN AUGMENTING CEPSTRAL FEATURES WITH FEATURES THAT SPAN MUCH LONGER STRETCHES OF TIME.

Similarly, new computational paradigms modeling multiple streams and their statistical dependencies come at increased computational and memory costs beyond those required by simple HMMs. These costs are higher because of the larger model size for coupled state and observation spaces, suggesting the need for faster probabilistic inference algorithms and judicious model selection methods for controlling model complexity.

STATISTICAL MODELING FOR THE NEW FEATURES

The results reported thus far were all obtained with HMM-based systems, which appear to have been adequate for the task of estimating and utilizing statistics from the new signal representations. In principle, however, HMMs are not well suited to long-term features. The HMM and frame-based cepstra have coevolved as ASR system components and hence are very much tuned to each other. The use of HMMs as the core acoustic modeling technology might obscure the gains from new features, especially those from long time scales; this may be one reason why progress with novel techniques has been so difficult. In particular, systems incorporating new signal processing methods in the front end are at a disadvantage when tested using standard HMMs. Additionally, the standard way to use longer temporal scales with an HMM is simply to use a large analysis window and a small (e.g., 10 ms) frame step, so that the frame rate is the same as for the small analysis window. The problem with this approach is that successive features at the slow time scale are even more correlated than those at the fast time scale, leading to a bias in posteriors. Models that do not represent the high correlation between successive frames effectively overcount the evidence about the underlying speech classes, creating a need for artificially weighting probability scores depending on the different time scales (i.e., longer time scales should have lower weight because of the higher correlation due to window overlap).

These points suggest that we should consider changing the statistical model. One approach that has been proposed is to add feature dependencies or explicitly model the dynamics of frame-based features in various extensions of HMMs [5], [24]. While we have made contributions in this area, we now believe that a very different approach is needed, one that relaxes the frame-based processing constraint. We propose instead to focus on the problem of multistream and multirate process modeling for two main reasons. First, it is desirable to improve robustness to corruption of individual streams. The use of multiple streams introduces more flexibility in characterizing speech at different time and frequency scales, which we hypothesize will be useful for both noise robustness and characterizing the variability observed in conversational speech. Second, the statistical models and features interact, and simple HMM-based combination approaches (or the other approaches that do not represent multiscale nature in multiple feature sequences) might not fully utilize complementary information

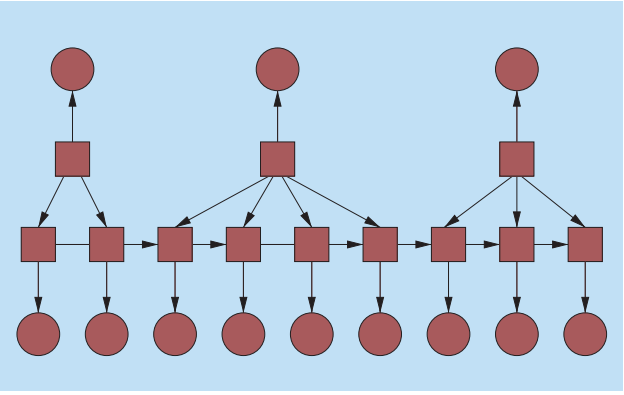


[FIG3] Streamlined training schedule for large MLP backpropagation learning task. A comparatively large learning rate (typically $\alpha = .001$) is used with a small fraction of the total training data (*N* was about 1,000 hours per gender) to begin the net training. The schedule parameters were extrapolated from many smaller experiments. Each of the four networks ({male,female}X{HATs,PLP Net}) has roughly eight million parameters and is trained on about 360 million speech feature vectors.

in different feature sequences, especially those from long time scales. In particular, we hypothesize that both the redundancy reduction and the selection of appropriate-sized modeling units are important for utilizing TRAPs- or HATs-like, long-term features in speech recognition.

We have experimented with an acoustic model using a multirate, coupled HMM architecture for incorporating acoustic and linguistic information from long time scales into speech recognition by joint statistical modeling of speech at phone and syllable time scales. In a two-rate, HMM-based acoustic model, we modeled speech using both recognition units and feature sequences corresponding to phone and syllable time scales. The short time scale in this model corresponds to the traditional phone HMMs using cepstral features, whereas the long time scale characterizes syllable structure and lexical stress using HATs. We have used the usual short-term spectral (PLP cepstral) features for the short-term scale modeling, whereas we used the long-term temporal ones (HATs) for the long-term scale modeling. Unlike the previously mentioned HAT features that were trained on phone targets, these HAT features are trained on broad consonant/vowel classes with distinction for syllable position (onset, coda, and ambi-syllabification) for consonants and low/high stress level for vowels. The HAT features trained on these latter targets provide much needed complementary information for multiscale modeling. The modeling units for the two time scales are phones and broad consonant/vowel phone classes. Words are represented as two parallel streams of context-dependent consonant/vowel states in the long time scale and as the usual context-dependent phone states in the short time scale. The alignment between two model sequences is allowed to vary across segment boundaries corresponding to syllable structure, allowing partial asynchrony in time and in modeling units.

WE PROPOSE TO FOCUS ON THE PROBLEM OF MULTISTREAM AND MULTIRATE PROCESS MODELING.



[FIG4] A graphical model illustration of a two-scale, variable-rate model unfolded in time, with the long time scale at the top. The squares represent hidden states associated with each scale, whereas the circles represent corresponding observations. The observations at the long time scale are irregularly sampled (as compared to regularly sampled short time scale observations) so as to focus more on temporally varying regions over the signal space.

In a medium-vocabulary version of the Switchboard conversational telephone speech recognition task, we have found that the explicit modeling of speech at two time scales via multirate, coupled HMMs architecture outperforms simple HMM-based feature concatenation and multistream modeling without downsampling approaches; these results emphasize the importance of redundancy reduction for knowledge integration and the importance of accommodating new features with new statistical models. The best performance, however, was obtained by a variable-rate extension in which the number of observations at one scale corresponding to an observation at the next coarsest scale can vary [8]. As such, the feature extraction and statistical modeling are tailored to focus more on information-bearing regions (e.g., phone transitions) as opposed to a uniform emphasis over the whole signal space. In these experiments, we dynamically sampled the coarse features (HATs) only when they significantly differed from the one occurring before so that, on average, one in three coarse features was kept. Figure 4 shows a graphical model illustration of a two-rate model using the variable-rate extension in aligning observations from different scales.

In our two-scale multi- and variable-rate models, we assumed that the coarser scale is about three times slower than the finer scale. The rate factor, three, was largely determined from minimum length constraints to cover speech utterances with left-to-right coarser-scale state topologies without skips. One interesting and potentially fruitful research direction is a more careful choice of this sampling rate, according to, for example, the scale/rate of the larger time-window features. Another interesting research direction is the multirate acoustic models with more than two time scales. The third or higher time scale can represent utterance-level effects such as speaking rate and style, gender, and noise.

WHAT COULD BE NEXT

The incorporation of temporal information via TRAPs or HATs appears to complement the more standard front end well enough to markedly improve recognition performance. However, this method is clearly a special case of a much broader class of front ends. In [22], a set of Gabor filters was applied to a log mel spectrogam, creating sensitivity to spectral, temporal, and spectrotemporal patterns at various scales. A subset of these filters was then chosen to optimize classification performance. This resulted in performance improvements for a small noise robustness task. This approach could be a candidate for future work in the search for a more general solution to the question of which combination of which feature streams would actually be optimal for

large speech recognition tasks. In a related technique, we are working to determine optimal window sizes and frame rates for different regions of speech, thus creating a signal-adaptive front end [27]. Incorporating such a subsystem can require significant changes to the rest of the system, and so our early work in this area is being conducted using rescoring approaches (e.g., rescoring a list of the N most probable word sequence hypotheses with probabilities arising from one or more alternate analysis windows).

The energy-based representations of temporal trajectories (or, more generally, of spectrotemporal patterns) could be replaced by autoregressive models for these components of the time-frequency plane, such as the FDLP or LP-TRAP approaches described earlier. We are also looking at combining the two dual forms of linear prediction, creating a new representation called perceptual linear prediction squared (PLP^2) [5]. PLP^2 is a spectrogram-like signal representation that is iteratively approximated by all-pole models applied sequentially in the time and frequency directions of the spectrotemporal pattern. First, the Hilbert envelopes in critical-band subbands are fit (so-called subband FDLP), and then all-pole models are fit to smooth the energy across subbands. Note, however, that unlike conventional feature processing, no frame-based spectral analysis occurs. After a few iterations, the result converges to a two-dimensional pattern that can be represented by a series of all-pole vectors in either the time or the frequency domain.

Finally, we note the related and impressive work done recently at IBM on incorporating low-dimensional projections of high-dimensional, Gaussian-derived posteriors, also incorporated in the feature vector used for recognition [25]. In this case also they seem to get significant benefit from posteriors incorporating information from long time spans.

FINAL WORDS

In [7] we whimsically referred to the increase in error rates as a goal in speech recognition research. Of course, we did not mean this literally; rather, we intended to encourage intrepid exploration of the road "less traveled" [15]. We implored the reader not to be deterred by initial results that were poorer than those achieved by more conventional methods, since this was almost inevitable when wandering from a well-worn path. However, the goal was always to ultimately improve performance, and the explorations into relatively uncharted territory were only a path to that goal. This process can be slow and sometimes frustrating. But eight years after we told this story, we are now seeing some of the improvements on large tasks that we earlier saw hints of in small tasks.

ACKNOWLEDGMENTS

This work was made possible by funding from the DARPA EARS "Novel Approaches" Grant MDA972-02-1-0024. We are also grateful to Jeff Bilmes for his work on GMTK, which was extensively used in the multirate/variable-rate experiments. The anonymous reviewers also made many useful suggestions, many of which we actually used. Finally, we thank the many other contributors in the speech groups at our five sites who contributed to our methods over the last three years.

AUTHORS

Nelson Morgan received his Ph.D. in electrical engineering from the University of California at Berkeley in 1980. He is the director of the International Computer Science Institute, where he's worked on speech processing since 1988. He is also a professor-in-residence with the Electrical Engineering Department of the University of California, Berkeley, and has over 175 publications, including three books. He is a Fellow of the IEEE.

Qifeng Zhu received his Ph.D. from the University of California, Los Angelos, in 2001. He is currently a senior researcher at the International Computer Science Institute. He was a research engineer at Nuance Communications from 2001–2002. His research focus includes speech signal processing, robust speech recognition, large-vocabulary speech recognition, and speech coding.

Andreas Stolcke received a Ph.D. in computer science from the University of California at Berkeley. He was an International Computer Science Institute postdoctoral researcher, working on probabilistic methods for natural language processing and learning. He is a senior research engineer at SRI International and the International Computer Science Institute, leading automatic speech and speaker recognition projects.

Kemal Sönmez is a senior research engineer in the STAR Lab at SRI International. He received his Ph.D. in electrical engineering from University of Maryland, College Park, in 1998. His research interests include statistical signal processing, speech and speaker recognition, genomic signal processing, and applications of bioinformatics in systems biology.

Sunil Sivadas received a M.Tech. degree from the Center for Electronic Design and Technology, I.I.Sc., India, in 1998 and a Ph.D. in electrical engineering from Oregon Health and Science University, Portland, in 2004. He is currently with Nokia Research Center, Tampere, Finland.

Takahiro Shinozaki received B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology, in 1999, 2001, and 2004, respectively. He is currently a research scholar in the Department of Electrical Engineering, University of Washington. His research interests include acoustic and language modeling for spontaneous speech recognition.

Mari Ostendorf received her Ph.D. in electrical engineering from Stanford University in 1985. She has worked at BBN Laboratories, Boston University, and is now a professor of electrical engineering at the University of Washington. She has over 150 publications in speech/language processing and is a Fellow of the IEEE.

Pratibha Jain received her Ph.D. from the OGI School of Science and Engineering, Oregon Health and Science University, in 1997 and her master's degree from IIT, Kanpur, India, in 2003. She joined STAR Labs, SRI, in 2003. Since 2004, she is with TI, Bangalore MMCODEC group. Her area of interest is robust speech processing.

Hynek Hermansky received the Dr.Eng. from the University of Tokyo and the Dipl.Ing. from Brno University of Technology, Czech Republic. He is with the IDIAP Research Institute, Martigny, Switzerland. He is a professor at the OGI School of Oregon Health and Science University and at Brno University of Technology. He has been an External Fellow at ICSI Berkeley and a visiting professor at the Chinese Academy of Sciences. He is a member of the board of ICSA and of the editorial boards of *Speech Communication* and *Phonetica*. He is a Fellow of the IEEE.

Dan Ellis is an assistant professor in the Electrical Engineering Department at Columbia University, where he heads the Laboratory for Recognition and Organization of Speech and Audio (LabROSA). His group investigates analysis of and information extraction from all kinds of audio, including speech, music, and everyday acoustic ambience.

George Doddington did his doctoral research on speaker recognition at Bell Labs and then spent 20 years directing speech research at Texas Instruments. He then spent ten years contributing to human language technology (HLT) at DARPA, NSA, and NIST. He is now an independent contractor supporting HLT research and evaluation.

Barry Chen received the B.S. degree in electrical engineering from the University of Maryland, College Park, in 1997 and the Ph.D. degree in electrical engineering from the University of California at Berkeley in 2005. His interests include neural networks, graphical models, and general systems for pattern recognition.

Özgür Çetin is a postdoctoral researcher at the International Computer Science Institute, Berkeley. He received a B.S. degree from Bilkent University, Turkey, in 1998 in electrical and electronics engineering. He has received the M.S. and Ph.D. degrees from the University of Washington, Seattle, in 2000 and 2005, respectively, both in electrical engineering.

Hervé Bourlard is director of the IDIAP Research Institute and a professor at the Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland. He is author/coauthor/editor of four books and over 190 reviewed papers and book chapters (including one IEEE paper award). He is a Fellow of the IEEE.

Marios Athineos received the B.S. degree in physics from the University of Patras, Greece, in 1998, and the M.S. degree in electrical engineering from Columbia University, New York, in 2000. He is currently a Ph.D. candidate in electrical engineering at Columbia University, conducting research on novel feature extraction methods.

REFERENCES

[1] J. Allen, "How do humans process and recognize speech?" *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[2] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[3] M. Athineos and D.P.W. Ellis, "Frequency-domain linear prediction for temporal features," in *Proc. ASRU*, 2003, pp. 261–266.

[4] M. Athineos, H. Hermansky, and D. Ellis, "LP-TRAP: Linear predictive temporal patterns," in *Proc. ICSLP*, 2004, pp. 949–952.

[5] M. Athineos, H. Hermansky, and D. Ellis, "PLP^2: Autoregressive modeling of auditory-like 2-D spectro-temporal patterns," in *Proc. ISCA Tutorial Research Workshop Statistical and Perceptual Audio Processing SAPA–04,* Jeju, Korea, Oct. 2004, pp. 37–42.

[6] J. Bilmes, "Maximum mutual information based reduction strategies for cross-correlation based joint distributional modeling," in *Proc. ICASSP-98*, Seattle, 1998, pp. 469–472.

[7] H. Bourlard, H. Hermansky, and N. Morgan, "Towards increasing speech recognition error rates," *Speech Commun.*, vol. 18, no. 3, pp. 205–231, May 1996.

[8] Ö. Çetin and M. Ostendorf, "Multi-rate and variable-rate modeling of speech at phone and syllable time scales," in *Proc. ICASSP 2005,* pp. I-665–668.

[9] B. Chen, Ö. Çetin, G. Doddington, D. Morgan, M. Ostendorf, T. Shinozaki, and Q. Zhu, "A CTS task for meaningful fast-turnaround experiments" in *Proc. RT-04 Workshop*, IBM Palisades Center, Nov. 2004.

[10] B. Chen, Q. Zhu, and N. Morgan, "Learning long term temporal features in LVCSR using neural networks," in *Proc. ICSLP,* 2004, pp. 612–615.

[11] E. Davis and O. Selfridge, "Eyes and ears for computers," *Proc. IRE,* vol. 50, pp. 1093–1101, 1962.

[12] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soci. Series B*, vol. 39, pp. 1–38, 1977.

[13] L. Deng and D. Sun, "Phonetic classification and recognition using HMM representation of overlapping articulatory features for all classes of English sounds," *Proc. ICASSP*, Apr. 1994, pp. 45–48.

[14] H. Dudley, "The vocoder," *Bell Labs Record*, vol. 17, pp. 122–126, Dec. 1939.

[15] R. Frost, "The road not taken," in *Mountain Interval*. New York: Henry Holt and Co., 1920.

[16] S. Furui, "Speaker independent isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. Acoust. Speech Audio Processing,* vol. 34, no. 1, pp. 52–59, 1986.

[17] R. Haeb-Umbach, D. Geller, and H. Ney, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, Adelaide, Australia, 1994, vol. 2, pp. 239–242.

[18] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing (Special Issue on Robust Speech Recognition)*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[19] H. Hermansky and S. Sharma, "TRAPS—Classifiers of temporal patterns," in *Proc. ICSLP-98*, Sydney, 1998, vol. 3, pp. 1003–1006.

[20] H. Hermansky, S. Sharma, and P. Jain, "Data-derived nonlinear mapping for feature extraction in HMM," in *Proc. ASRU-99*, Keystone, CO, 1999, pp. I-63–66.

[21] M. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proc. IEEE Conf. Acoustics, Speech, Signal Processing*, Glasgow, Scotland, 1989, pp. 262–265.

[22] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," in *Proc. ICSLP-2002*, Denver, CO, Sept. 2002, pp. 25–28.

[23] S. Lee and J. Glass, "Real-time probabilistic segmentation for segment-based speech recognition," in *Proc. ICSLP-1998*, Sydney, 1998, pp. 1803–1806.

[24] M. Ostendorf, V. Digilakis, and O. Kimball, "From HMMs to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 4, no. 5, pp. 369–378, 1996.

[25] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "FPME: Discriminatively trained features for speech recognition," in *Proc. RT-04 Workshop*, IBM Palisades Center, Nov. 2004.

[26] A. Robinson, M. Hochberg, and S. Renals, "IPA: Improved modelling with recurrent neural networks," in *Proc. ICASSP-94*, Apr. 1994, pp. 37–40.

[27] M. Sonmez, M. Plauche, E. Shriberg, and H. Franco, "Consonant discrimination in elicited and spontaneous speech: A case for signal-adaptive front ends in ASR," in *Proc. ICSLP-2000*, Beijing, China, Oct. 2000, pp. 548–551.

[28] H. Yang, S. Van Vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Commun.*, vol. 31, no. 1, pp. 35–50, 2000.

SP