TRANSITION-BASED STATISTICAL TRAINING FOR ASR

Nelson Morgan [‡,*], *Yochai Konig* [‡,*], *Su-Lin Wu* [‡,*], *and Hervé Bourlard* [‡,†],

[‡] International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704
Phone: (510)-643-9153 Fax: (510)-643-7684
Email: morgan, konig, sulin, bourlard@icsi.berkeley.edu
(also [†] Faculté Polytechnique de Mons, Belgium and [*] University of California at Berkeley)

1. INTRODUCTION

It is known that in human speech recognition, the perceptually-dominant and information-rich portions of the speech signal, which may also be the parts with a better chance to withstand adverse acoustical conditions, are the (phonetic) transitions (see, e.g., [3] for some experimental evidence).

A first step in this direction was to use highpass or bandpass filtering of critical band trajectories (RASTA processing) to emphasize transitions [5]. While this is sometimes helpful in reducing errors due to (channel) mismatches between training and testing conditions, the resulting observation sequence is a representation that has emphasized the regions of strong change and de-emphasized temporal regions without significant spectral change. This is a mismatch to the underlying speech model in standard HMMs, which has been designed to represent piecewise stationary signals. In general, modeling transitions or any non-stationary properties of speech signal require major modifications of standard HMMs [4]. Therefore, it is likely that transition-based systems will require a fundamentally different kind of underlying statistical model. We have been developing a statistical model (SPAM) and a statistical training algorithm (REMAP) that may be more appropriate to this perspective.

2. SPAM: STOCHASTIC PERCEPTUAL AUDITORY-EVENT-BASED MODELS

Speech can be viewed as a sequence of Auditory Events (Avents), which are elementary decisions made in response to significant changes in spectral amplitudes (as in [3]). Avents are presumed to occur about once per phone boundary. The statistical model uses these Avents as fundamental building blocks for words and utterances, separated by states corresponding to the more stationary regions. In order to focus the statistical power on the rapidly-changing portions of the time series, all of the non-Avent states are tied to the same class. Markov-like recognition models use Avents as time-asynchronous observations. Discriminant models are trained to distinguish among all classes (including the non-Avent class).

In this case, SPAMs are defined from a set of Avents $\mathcal{Q} = \{q^0, q^1, \ldots, q^K\}$, in which all q^k's, for $k \neq 0$, represent Avents and q^0 represents the non-Avent or non-perceiving state. This set is currently initialized to correspond to left context-dependent phonetic onsets.

As discussed in [6], one can do SPAM recognition based on the following local acoustic probabilities:

$$P(q_n^\ell | q_{n-\Delta(n)}^k, \Delta(n), X_{n-d}^{n+c}), \quad \left\{ \begin{array}{l} \forall \ell = 0, 1, \ldots, K \\ \forall k = 1, 2, \ldots, K \end{array} \right\} \quad (1)$$

in which $X_{n-c}^{n+d} = \{x_{n-c}, \ldots, x_n, \ldots, x_{n+d}\}$ and $n - \Delta(n)$ corresponds to the previous time index for which an Avent had been perceived and becomes one of the stochastic variable of the model. During training, these local probabilities are estimated by an artificial neural network (ANN), via an iterative Viterbi-like segmentation to provide the net with targets or via REMAP (as explained in Section 3). According to our SPAM constraints, these local probabilities are used for training and decoding in particular left-to-right HMMs constituted by sequences of Avent states (with no loop allowed) separated by (looped) tied non-Avent states.

Preliminary experiments on isolated telephone digits (plus yes and no) are reported in [6]. In these experiments, in which we simplify (1) to the simple Avent posterior $P(q_n^\ell | X_{n-d}^{n+c})$, it is shown that the reduced SPAM has about double the error rate of our best phone-based system. However, in another experiment in which we artificially reduced the SNR to 10 dB by adding car noise, the error rates of the two systems were comparable. Combination of both approaches (using a weighted sum of word scores) led to similar performance for the clean case but to a 30% relative reduction in errors for the noisy case (reduction from 10.9% error to 7.7% on 2600 examples from 200 speakers, using 4 jackknifed testing cuts, where training was always with clean data).

3. REMAP

Our work with transition-based models motivated us to develop training algorithms that are more appropriate to these models. In particular transition-based models require a smooth estimate of transition probabilities and also tend to use posterior probabilities (unlike the scaled likelihoods in our standard system). To this end, we have developed the REMAP (Recursive Estimation and Maximization of A Posteriori Probabilities) learning algorithm.

Estimating transitions accurately is difficult. In our previous hybrid systems, the targets used for ANN training are typically given by the best segmentation resulting from a Viterbi alignment. This procedure thus yields rigid transition targets, which may not be optimal in the case of training (and testing!) of posterior probabilities for SPAMs. Additionally, other work we are doing in

transition-based systems requires phone posteriors conditioned on the previous state, and this too requires the identification of transitions (at least implicitly). A better goal might be to learn smooth probabilities for phonetic transitions conditioned on the acoustics.

Additionally, our training algorithm should be based on learning posteriors that are local (in time) such that we directly optimize the parameter set Θ according to the MAXIMUM A POSTERIORI (MAP) criterion, i.e., maximizing $P(M|X,\Theta)$ if M is the correct HMM associated with acoustic data X. In theory such an optimization would minimize the utterance error rate. REMAP successively estimates new (local) posterior probabilities to be used as targets for ANN training, guaranteeing an iterative increase of the global posteriors. Estimation of the new network targets requires "forward" and "backward" recurrences that are reminiscent of the EXPECTATION MAXIMIZATION (EM) algorithm.

3.1. PROBLEM FORMULATION

We wish to find the optimal parameter set Θ maximizing

$$\prod_{i=1}^{I} P(M_i|X_i,\Theta) \qquad (2)$$

in which M_i represents the Markov model associated with each training utterance X_i, with $i = 1,\ldots,I$.

EM-like MAP training of transition-based HMM/ANN hybrids requires a solution to the following problem: given a trained ANN at iteration t providing a parameter set Θ^t and, consequently, estimates of $P(q_n^\ell|x_n,q_{n-1}^k,\Theta^t)$, how can we determine new ANN targets that:

1. Will be smooth estimates of conditional transition probabilities, $\forall$ possible (k,ℓ) state transition pairs in M and $\forall n \in [1,n]$.
2. When training the ANN for iteration $t+1$, will lead to new estimates of Θ^{t+1} and $P(q_n^\ell|x_n,q_{n-1}^k,\Theta^{t+1})$ that are guaranteed to incrementally increase (2)?

In [2], we prove that a re-estimate of ANN targets that guarantee convergence to a local maximum of (2) is given by:

$$P^*(q_n^\ell|x_n,q_{n-1}^k) = P(q_n^\ell|X,q_{n-1}^k,\Theta^t,M) \qquad (3)$$

which means that the new ANN target associated with x_n and a specific transition $q^k \rightarrow q^\ell$ has to be calculated as the probability of that specific transition CONDITIONED ON THE WHOLE TRAINING SENTENCE X and the associated model M.

In [2], we further prove that alternating ANN target estimation (the "estimation" step) and ANN training (the "maximization" step) is guaranteed to incrementally increase (2) over t; we also provide efficient forward and backward-like recurrences to compute (3).

3.2. DISCUSSION AND RESULTS

Of course, a wide range of discriminant approaches (e.g., MMI, GPD – see [2] for a discussion of these) to speech recognition have been studied by researchers. A significant difficulty that has remained in applying these approaches to continuous speech recognition has been the requirement to run computationally intensive algorithms on all of the rival sentences. Since this is not generally feasible, compromises must always be made in practice. For instance, estimates for all rival sentences can be derived from a list of the "N-best" utterance hypotheses, or by using an ergodic model containing all possible phonemes. This is not required with the present algorithm.

Although much work is still required to optimize the practical heuristics for this method, preliminary results show a 27% relative reduction in error on telephone isolated digits (reduction from 3.4% error to 2.5% error for a case with similar car noise in both training and test).

4. CONCLUSIONS

We now have some theory and some initial results. We are currently working on including the dependence on time to previous Avent in the SPAM process, and are beginning to apply REMAP to continuous speech. We also have yet to explore the possible symbiosis between these approaches, although this has been implicit in our thinking over the last year. In this regard, we are exploring the use of acoustic transition probabilities from REMAP for training of SPAMs, instead of hard targets for phonetic onsets.

5. ACKNOWLEDGMENTS

Thanks to Steven Greenberg (from ICSI) and Hynek Hermansky (from OGI) for many useful discussions. The work was partly sponsored by the Joint Services Electronics Program (JSEP) Contract No. F49620-94-C-0038, and the Office Naval Research, URI Grant no. N00014-92-J-1617 (via UCB), and ESPRIT project 6487 (WERNICKE) (through ICSI). Su-Lin Wu was partially supported by a National Science Foundation Fellowship.

6. REFERENCES

[1] Bengio, Y. R., De Mori, R., Flammia, G., & Kompe, R. (1992). "Global optimization of a neural-hidden Markov model hybrid," IEEE TRANS. ON NEURAL NETWORKS, vol. 3, pp. 252-258.

[2] Bourlard, H., Konig, Y., & Morgan, N., "REMAP: recursive estimation and maximization of a posteriori probabilities – Application to transition-based connectionist speech recognition," ICSI TECHNICAL REPORT TR-94-064, INTL. COMPUTER SCIENCE INSTITUTE, BERKELEY, CA, 1994.

[3] Furui, S., "On the role of spectral transition for speech perception," J. ACOUST. SOC. AM., vol. 80, no. 4, pp. 1016-1025, 1986.

[4] Ghitza, O. and Sondhi, M.M., "Hidden Markov models with templates as non-stationary states: an application to speech recognition," COMPUTER SPEECH AND LANGUAGE, 2:101–119, 1993.

[5] Hermansky, H. and Morgan, N., "RASTA processing of speech", IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, special issue on Robust Speech Recognition, vol.2 no. 4, pp. 578-589, Oct. 1994.

[6] Morgan, N., Wu, S.-L., & Bourlard, H., "Digit recognition with stochastic perceptual models," to be published in PROC. EUROSPEECH'95 (Madrid, Spain), September 1995.