LANDMARK-BASED SPEECH RECOGNITION: REPORT OF THE 2004 JOHNS HOPKINS SUMMER WORKSHOP

Mark Hasegawa-Johnson [1], *James Baker* [2], *Sarah Borys* [1], *Ken Chen* [1]
Emily Coogan [1], *Steven Greenberg* [3], *Amit Juneja* [4], *Katrin Kirchhoff* [5], *Karen Livescu* [6]
Srividya Mohan [7], *Jennifer Muller* [8], *Kemal Sonmez* [9], *Tianyu Wang* [10]

1. University of Illinois, 2. Carnegie Mellon, 3. University of California,
4. University of Maryland, 5. University of Washington, 6. MIT,
7. Johns Hopkins, 8. Department of Defense, 9. SRI, 10. Georgia Tech

ABSTRACT

Three research prototype speech recognition systems are described, all of which use recently developed methods from artificial intelligence (specifically support vector machines, dynamic Bayesian networks, and maximum entropy classification) in order to implement, in the form of an automatic speech recognizer, current theories of human speech perception and phonology (specifically landmark-based speech perception, nonlinear phonology, and articulatory phonology). All three systems begin with a high-dimensional multiframe acoustic-to-distinctive feature transformation, implemented using support vector machines trained to detect and classify acoustic phonetic landmarks. Distinctive feature probabilities estimated by the support vector machines are then integrated using one of three pronunciation models: a dynamic programming algorithm that assumes canonical pronunciation of each word, a dynamic Bayesian network implementation of articulatory phonology, or a discriminative pronunciation model trained using the methods of maximum entropy classification. Log probability scores computed by these models are then combined, using log-linear combination, with other word scores available in the lattice output of a first-pass recognizer, and the resulting combination score is used to compute a second-pass speech recognition output.

1. INTRODUCTION

Humans and machines recognize consonants on the basis of acoustic cues present just after consonant release, and just before consonant closure; acoustic spectra during the closure interval itself provide little phonetic information [1]. Stevens has proposed [2] that consonant closures and releases, as well as syllable peaks and dips, compose a series of "acoustic landmarks" around which human and automatic speech recognition may be organized. Detection of these landmarks provides two sets of cues to a human or automatic speech recognizer: (1) detected manner-change landmarks specify the manner of articulation (stop, nasal, fricative, glide, vowel) of the phonemes, and (2) manner-change landmarks can be used to synchronize classifiers that seek to identify place and voicing.

Conversational telephone speech is characterized by variable pronunciation. Many common pronunciation variants are generated by the reduction, overlap, and assimilation of distinctive features, and can thus be compactly represented in a landmark-based speech recognition model. Distinctive-feature-based models may be divided into roughly two categories, depending on whether the structure they impose is primarily production-oriented or primarily perception-oriented. The theory of articulatory phonology [3] is primarily production-oriented: pronunciation variability is modeled as asynchrony and overlap among articulatory gestures, and is constrained by the structure of the vocal tract. Stevens et al. [2] proposed a model that could be considered perception-oriented: in their model, the distinctive features in each word are either "modifiable" or "required," where the difference between these two categories is partly determined by the degree to which each feature disambiguates the word from phonemically similar or confusable words.

This paper describes three probabilistic landmark-based speech recognizers, developed at the workshop (WS04) at the Johns Hopkins Center for Language and Speech Processing. In all three systems, a high-dimensional acoustic feature vector is transformed into a vector of posterior probability estimates by support vector machines (SVMs) trained to detect landmarks, and to classify distinctive features. The first system uses a dynamic programming algorithm to combine SVM-computed acoustic probabilities, in order to estimate the probability that a word has been produced using its canonical dictionary pronunciation. The second system uses a dynamic Bayesian network to represent the asynchronous, independently reduced articulatory gestures posited by articulatory phonology. The third system uses the maximum entropy method (ME) to determine which of the distinctive features in each word are required to disambiguate speech recognizer output, and then computes a word score by considering SVM outputs corresponding only to the required distinctive features.

SVMs are trained and tested using NTIMIT and the

This research would have been impossible without the support of Andreas Stolcke, John Makhoul, Owen Kimball, Spyros Matsoukas, Dimitra Vergyri, Luciana Ferrer, Nima Mesgarani, Yanli Zheng, Tarun Pruthi, Shihab Shamma, Carol Espy-Wilson, Jeff Bilmes, Partha Niyogi, Jim Glass, T.J. Hazen, Eric Fosler-Lussier, Fred Jelinek, Sanjeev Khudanpur, the National Science Foundation, and the Department of Defense.

phonetically transcribed portion of Switchboard. Half of the talkers in the transcribed Switchboard set are used to train SVMs, and half to test. Rescoring systems are tested using the development test data from the 2003 NIST rich text transcription task (RT03). The evaluation task for pronunciation models is lattice rescoring using word lattices generated by the SRI speech recognizer; word lattices for the Hub-5 training and evaluation speech data were also provided by BBN. Pronunciation model output probabilities are combined with the HMM-based acoustic model scores and N-gram based language model scores using a log-linear combination algorithm. Different experiments used three different methods to set the scaling coefficients for log-linear combination: scaling coefficients are either set heuristically, or using amoeba search [4], or using a novel discriminative exponential model designed to minimize word error rate. Similarly to the approach proposed in [5], the exponential model is conditioned on context via a set of features, whose weights are estimated using ME. The model tested here differs from that proposed in [5] in that it targets word error rate (WER) by working within the framework of confusion networks, a compact representation of hypotheses in the lattices as described in [6]. In the algorithm proposed here, log linear weight estimation is posed as ME estimation of the conditional exponential model for the probability that a hypothesized word in a confusion network is the reference. Features to represent the confusion network context included normalized posterior rank (rank/#words), original posterior, landmark pronunciation model scores (DBN scores, discriminative pronunciation model scores), original acoustic and language model scores, duration, number of phones, relative confusion network position in the lattice, confusability penalty, and function word set membership.

Accuracy of the SVMs improved steadily throughout the workshop. Word error rate (WER) reductions were achieved on training data, and on a three-speaker subset of the development test data, but no statistically significant WER reductions were achieved on the complete RT03 development test set.

2. LANDMARK DETECTION AND CLASSIFICATION

All systems described in this paper observe a composite acoustic feature vector including the following features: energy, spectral tilt, spectral compactness, MFCC, formant frequency, amplitude, and bandwidth [7], knowledge-based acoustic features designed to be informative about the phonological distinctive features [8], and the "rate-scale" auditory cortical features [9]. All of these features are measured once per 5 ms, except that energy, spectral tilt, and spectral compactness are measured once per millisecond. Each SVM is trained to compute a real-valued linear or nonlinear discriminant function, $g_j(X_t)$, where j is the index of the distinctive feature, and X_t is the concatenation of 4 to 17 acoustic feature frames sampled in the vicinity of frame t. The discriminants are mapped to pseudo-posteriors using a histogram. Histogram counts are trained using a corpus with equal numbers of positive and negative examples ($q(d_j = +1) = 0.5$), so that the pseudo-posterior $q(d_j|g_j(X))$ is proportional to the true likelihood $p(g_j(X)|d_j)$. Posterior probabilities of manner features are computed independently of the settings of all other distinctive features. Posterior probabilities of place and voicing features are computed using context-dependent SVMs, meaning that a bank of SVMs and a corresponding bank of histograms are trained to estimate $p(d_j(t) = 1|X_t, L(t))$ for every possible value of $L(t)$, where $L(t)$ specifies that t is a landmark of a particular type (stop release, stop closure, fricative release, etcetera).

Several hundred SVMs were trained, using linear and RBF kernels; 72 were selected for use in the pronunciation models. Acoustic feature inputs were selected separately for each of the 72 classifiers, but generally included between 3 and 30 acoustic features for manner classification, and between 500 and 2000 acoustic features for place or voicing classification. It was discovered that place of articulation classifiers typically improve with every increase in the acoustic feature vector dimension, provided that the new features are not completely determined by existing features, and provided that the dimension of the acoustic feature vector does not exceed roughly 17% of the number of frames in the training corpus. Prior to the start of the workshop, classification error rates of most manner features were already at a very low level [10], but classification error rates of most place and voicing features were inadequate for speech recognition purposes. During the course of the workshop, classification error rates of all place and voicing features fell by 10-50%, through a combination of improved selection of acoustic features and improved classifier training methods. Error of nasal place classifiers, for example, fell by 49%; error rate of stop place classifiers fell by 21%.

The landmark-based recognition paradigm allowed us to explore phonetic distinctions that are ignored by most English-language speech recognizers. In the Switchboard transcriptions, for example, nasalized vowels were often found in place of deleted nasal phonemes. We reasoned, therefore, that the pronunciation model should be given the ability to learn that a nasalized vowel is a high-probability substitute for a nasal consonant, and that therefore, it would be useful to develop a detector for nasalized vowels. Two types of nasalized vowel detectors were developed: a classifier that distinguished all nasalized vowels from all non-nasalized vowels, and a bank of vowel-dependent nasalization detectors. Vowel nasalization proved difficult to classify accurately, but reasonably successful classifiers were trained for four vowels: /ey/ (81% accuracy, in a test set with 50% nasal vowels), /iy/ (76%), /ae/ (75%), and /ao/ (73%).

Landmarks were detected by a dynamic programming algorithm that combines information about manner class observation probabilities, together with a manner-class probability model, in order to reduce the number of false landmark insertions [10]. SVM-based classifiers for the manner features were applied in each frame of speech. SVM-computed probabilities were combined with a segmentation algorithm to obtain the manner change landmarks - fricative closure and release, sonorant consonant closure and release, vowel nucleus, syllabic dip, silence start and end, stop burst and vowel onset.

The dynamic programming algorithm is itself a speech recognizer, and was used in lattice rescoring experiments. Probability of a word in this method is computed as $P(U|O) =$

$P(L|O)P(U|LO)$, where U is a sequence of bundles of distinctive features or the corresponding sequence of phones, L is the canonical sequence of landmarks and O is the sequence of all acoustic observations. Good phoneme alignment was achieved using this method, but no WER reduction was achieved, apparently because few words in the conversational speech corpus are adequately modeled by their canonical landmark sequences.

3. RESCORING USING A GENERATIVE FEATURE-BASED PRONUNCIATION MODEL

A hybrid SVM-DBN landmark-based speech recognizer was created by combining the generative pronunciation model of [11] with the SVM acoustic observation probabilities described in Sec. 2. In the generative pronunciation model, hidden variables in a DBN represent features based on the tract variables of [3], including the locations and/or degrees of opening of the lips, tongue tip, and tongue body, and the states of the glottis and velum. Pronunciations are generated by mapping a word's baseform pronunciations to trajectories, allowing the tract variables to go through their trajectories asynchronously (while enforcing some soft synchrony constraints, encoded as distributions over degrees of asynchrony), and allowing each feature's surface value to stray from its underlying target value (typically due to undershoot) with some probability. The use of a DBN allows us to take advantage of the natural factorization of the large state space.

In order to incorporate the likelihoods from the SVMs, we used the Bayesian network construct of *soft evidence*. For each distinctive feature d_j, a "dummy" variable $\hat{g}_j$ is created, whose value is always 1 and whose distribution is constructed so that $P(\hat{g}_j = 1|d_j)$ is proportional to the acoustic likelihood $p(g_j(X)|d_j)$ computed by the SVM. Since the pronunciation model uses a different set of features from the SVMs, we used a deterministic mapping from articulatory to distinctive features, e.g., `sonorant` $= 1$ whenever the glottis is in the voiced state and either the lip and tongue openings are narrow or wider or there is a complete lip/tongue closure along with an open velum.

Since place and voicing SVMs are trained only in certain contexts ($p(d_j|X_t, L)$ is trained only using frames aligned with landmarks of type L), it is nonsensical to use all SVMs in all frames. For now, our solution is to rescore in two passes: The manner probabilities, which are interpretable in all frames, are used to obtain a manner segmentation; the DBN is then used along with the remaining SVM outputs to compute a score conditioned on the manner segmentation, using each SVM only in its appropriate context.

As a way of qualitatively examining the model's behavior, we can compute a "forced alignment" for a given waveform, i.e. the most probable values of all of the DBN variables given the word identities and the SVM outputs. Observation of forced alignments was used to correct problems in the integration of SVM scores into the DBN, and to evaluate the success of pronunciation modeling. It was observed that the system often produces the intuitively correct explanation for pronunciation reduction phenomena. For example, in one reduced utterance of the phrase "I don't know", the system was able to correctly recognize that both the /d/ and the /n t n/ sequence were produced essentially as glides, i.e., although the underlying setting of the tongue tip variable is `TT-OPEN=CL` ("closed"), the observed value is `ActualTT-OPEN=NA` ("narrow").

System setup	WER
Baseline	27.7
SVM-EBS-DBN	27.3
SVM-DBN-DBN	27.3
SVM-DBN-DBN, high-accuracy SVMs only	27.2

Table 1. Word error rates (%) in lattice rescoring experiments on a three-speaker (1988-word) subset of the RT03 development set. SVM-EBS-DBN refers to the case in which the event-based system (EBS) of [10] is used to do the manner segmentation; in the SVM-DBN-DBN case, the manner segmentation was done by the DBN using only the manner distinctive features.

Table 1 shows a sample of the word error rates obtained with this system on a three-speaker subset of the RT03 development set. All experiments were done using GMTK [12]. Because of time constraints during the workshop, we have not yet run all of the variants of the system on the full development or evaluation sets. For the only case in which we have a full development set result (the SVM-EBS-DBN case), there is no change from the baseline WER.

4. DISCRIMINATIVE RESCORING USING LANDMARKS

As an alternative to the generative pronunciation model, this section describes a model in which landmark information explicitly discriminates between confusable words in the baseline word lattice. The word lattice is first converted into a confusion network, as described in [6]. For each confusion set, words are then converted into a landmark-based representation suitable for training a maximum entropy (ME) discriminative classifier. The ME model requires a fixed-length input vector. Since words have different numbers of landmarks, we achieve a fixed-length representation by encoding them in terms of the frequencies of binary precedence and overlap relations between landmarks, e.g., "vowel" precedes "sonorant consonant" ($V \prec SC$), and "sonorant consonant" overlaps with "+blade" ($SC \circ +blade$). Not all possible precedence and overlap relations actually occur; in practice, the total number of relations is 40-60, depending on the specific set of landmarks used. The frequency of each relation within a word is entered into the respective element of the vector; the entire vocabulary can thus be represented as a matrix, similar to the word-frequency encoding of documents in Information Retrieval.

An ME model is trained to distinguish among the binary relationship vectors that correspond to the words in a particular confusion set. The features in the ME model are the landmarks relations described above. Ideally, vectors should be derived from a large training set consisting of time-aligned word and landmark transcriptions. Since such a training set was not available to us we used the word entries in a landmark-based pronunciation dictionary

sneak	speak
SC $\circ$ +blade 2.47	SC $\circ$ +blade -2.47
FR $\circ$ SC 2.47	FR $\prec$ SC -2.47
FR $\prec$ SIL -2.11	FR $\prec$ SIL 2.11
SIL $\prec$ ST -1.75	SIL $\prec$ ST 1.75
.....	

Fig. 1. Example of landmark weights to distinguish between *sneak* and *speak*. The highest weights are assigned to *sonorant consonant* overlapping with *+blade* (indicating the nasal /n/), and to *fricative* preceding *sonorant consonant*, indicating the /sn/ sequence.

as training samples. This dictionary (converted from an initial phone-based representation) includes the pronunciation variants used by the first-pass system, and uses a number of phonetic rules to derive a fine-grained landmark-based representation of pronunciation variants.

The trained maximum-entropy model assigns weights to each landmarks relation; the landmarks are then ranked according to the magnitude of each weight and the top N of these are selected. These are then passed back to the landmark detection module, together with the time boundaries of the words in question. The detection module performs a search for these landmarks within the time constraints specified and returns their combined log-likelihood. The weights of a trained model to distinguish between *sneak* and *speak* are shown in Figure 1.

Experiments were carried out on the RT03 development set. The baseline word error rate, obtained by selecting the highest scoring hypothesis in each confusion network, was 24.1%. Oracle word error rate varied between 22.0% and 23.9%, depending on the number of words retained in each confusion set (22.0% oracle WER was achieved by trimming each confusion set to five words, and counting homophones as correct; 23.9% oracle WER was achieved by trimming each confusion set to two words). Rescoring was done by a weighted combination of the baseline posterior probabilities and the normalized acoustic landmarks scores (weights 0.8 and 0.2, respectively). A number of variants of this technique were tested, including different weighting schemes for broad classes vs. place features, and a number of different methods for selecting landmark scores. Several methods resulted in a small reduction in the total number of word errors, but this number never totalled more than 0.05% of the WER denominator.

5. CONCLUSIONS

Methods described in this paper have resulted in WER reductions on an arbitrarily selected three-speaker subset of the target corpus, but no method applied to the entire corpus has resulted in a statistically significant WER reduction. Despite the current lack of a WER reduction, several intermediate evaluation results support the argument in favor of further research along these lines. Rapid and continuous gains in phonetic classification accuracy were achieved, relative to the start of the workshop. The SVM was proven capable of learning classification boundaries in a very high-dimensional observation space, typically on the order of 500 to 2000 observation dimensions. The DBN was shown capable of incorporating soft evidence computed by an SVM, and of using the available soft evidence to correctly transcribe consonant reductions.

Automatic classification of acoustic landmarks requires an algorithm capable of learning classification boundaries in a high-dimensional observation space; SVMs satisfy the requirement. Probabilistic modeling of articulatory asynchrony requires an algorithm capable of learning the joint distributions of many simultaneous hidden variables; DBNs satisfy the requirement. This paper has demonstrated that it is possible to build an automatic speech recognizer that learns, from data, some of the information structures apparently used in human speech perception and speech production.

6. REFERENCES

[1] Sadaoki Furui, "On the role of spectral transition for speech perception," *JASA*, vol. 80, no. 4, pp. 1016–1025, 1983.

[2] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a model for lexical access based on features," in *Proc. ICSLP*, Banff, Alberta, 1992, vol. 1, pp. 499–502.

[3] Catherine P. Browman and Louis Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[4] Dimitra Vergyri, "Use of word level side information to improve speech recognition," in *Proc. ICASSP*, 2000.

[5] H. Yu and A. Waibel, "Integrating thumbnail features for speech recognition using conditional exponential models," in *Proc. ICASSP*, 2004, pp. 893–896.

[6] Fuliang Weng, Andreas Stolcke, and Ananth Sankar, "Efficient lattice representation and generation," in *Proc. ICSLP*, 1998, pp. 2531–2534.

[7] Yanli Zheng and Mark Hasegawa-Johnson, "Formant tracking by mixture state particle filter," in *Proc. ICASSP*, 2004.

[8] Nabil Bitar and Carol Espy-Wilson, "A knowledge-based signal representation for speech recognition," in *Proc. ICASSP*, 1996, pp. 29–32.

[9] Nima Mesgarani, Malcolm Slaney, and Shihab A. Shamma, "Speech discrimination based on multiscale spectrotemporal features," in *Proc. ICASSP*, 2004.

[10] A. Juneja and C. Espy-Wilson, "Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition," in *From sound to sense: 50+ years of discoveries in speech communication*, MIT, Cambridge MA, 2004, pp. C–151 to C–156.

[11] Karen Livescu and James Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities," in *ICSLP*, 2004.

[12] Jeff Bilmes and Geoffrey Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," in *Proc. ICASSP*, 2002.