LATENT PHONETIC ANALYSIS: USE OF SINGULAR VALUE DECOMPOSITION TO DETERMINE FEATURES FOR CRF PHONE RECOGNITION

I.B. Heintz, E. Fosler-Lussier, C. Brew

The Ohio State University
Department of Linguistics & Department of Computer Science and Engineering
Columbus, OH 43204
{bromberg,cbrew}@ling.ohio-state.edu, fosler@cse.ohio-state.edu

ABSTRACT

We exploit an analogy between document retrieval and phone recognition, and adapt the method of Latent Semantic Analysis for the latter task. By mapping into a space of reduced dimensionality, we hope to uncover previously unexploited relationships between posterior estimates of phonetic events and the parts of phones represented by HMM states. We find that features defined over the reduced space complement those previously known, such as, for example, phonological features. We are able to effectively combine all of these features in a phone recognition task by using the constraint-based framework of Conditional Random Fields (CRFs), which allows the use of large and highly redundant feature spaces.

Index Terms— Speech recognition, Matrix decomposition, Stochastic fields

1. INTRODUCTION

In the task of phone recognition, we look for the differences and similarities among phones in order to best discriminate between them. We use our prior intuitions and knowledge about language and speech sounds to inform our exploration; we know that phonemes can be described in terms of phonological features like consonant voicing, vowel height, tenseness, etc. We (and others) have exploited this information in studies such as [1] and [2]. In this study, we aim to capitalize on relationships between phones that go beyond those that are designed into the phonological features. For instance, certain phones are prone to particular patterns of confusion, insertion or deletion, and the classes involved in these patterns do not necessarily align with the equivalence classes induced by the feature system. Specifically, we use a form of Latent Semantic Analysis, re-labeled as Latent Phonetic Analysis, to find relationships between phones that do not emerge when using typical automatic speech recognition (ASR) techniques like neural networks. We use this analysis to find generalizations about the phone classes and the acoustic data, and use these generalizations in consort with the relationships that are designed into the feature system. We slightly improve the phone recognition results by adding new features based on LPA, so we have reason to believe that we have captured phonetic information that goes beyond what was previously modeled.

This work was supported by NSF ITR grant IIS-0427413, NSF Career Grant 0643901, and by a student-faculty fellowship from the AFRL/Dayton Area Graduate Studies Institute. The opinions and conclusions expressed in this work are those of the authors and not of any funding agency.

1.1. Singular Value Decomposition

Acoustic modeling is a task in which phones or phonological features (objects) are discriminated on the basis of acoustic data (a much larger set of parameters). A similar task is the document modeling performed within IR, in which documents (objects) are discriminated on the basis of the many words (features) that comprise them. One method of document modeling is called Latent Semantic Analysis, or LSA ([3], [4]). LSA uses matrix decomposition to compare documents to each other along dimensions that reflect not only the words that comprise the individual documents, but also the relationships between the words. A goal of LSA is the identification and exploitation of synonymy relations. These relationships emerge naturally from a process of dimensionality reduction based on singular value decomposition (SVD). We aim to uncover analogous "synonymy" relations in the acoustic data that we use to model phonemes, so we have experimented with using the techniques of LSA to perform LPA - Latent Phonetic Analysis. There is ample precedent for this kind of re-use of LSA, as in [5].

We begin LPA by creating a phone-by-feature matrix with a cell for each potential association between a phone and a phonetic feature (described fully in Section 2). We then perform SVD over this matrix. SVD is the process by which an $N \times M$ matrix is decomposed into three component matrices:

$$A_{NxM} = U_{NxN} S_{NxM} V^T_{MxM} \quad (1)$$

Dimensionality reduction is performed by removing X columns from each matrix, those representing the lowest eigenvalues

and eigenvectors, such that:

$$A_{NxM} \approx U'_{NxY} S'_{YxY} V'^{T}_{YxM} \quad (2)$$

where $Y = M - X$ [6]. The rows of matrix U', with reduced dimensionality Y, model the phonetic features' relationships to each other. Features that are related in useful ways will, it is hoped, be near each other in the vector space described by all of the rows of U'. Similarly for V'^{T}, in which the columns are phone vectors that may be compared in vector space. SVD is designed to order the dimensions by importance, with the earlier dimensions accounting for the greatest variation. By removing the dimensions that correspond to smaller amounts of variation, we arrive at a compact representation of phones that will hopefully improve recognition by reducing noise and making evident the relevant relationships that were implicitly present in the original phone-by-feature matrix.

Our claim is that the SVD representation contains information that is not conveniently present in representations based more directly on phonetic features. We use this new information about phone and feature relationships to refine and expand our original acoustic models.

1.2. Conditional Random Fields

CRFs, as described in [7], discriminate between hypothesized label sequences. The hypotheses are conditioned on a set of arbitrary input features. In particular, the modeler is free to provide **overlapping** features, and has no need to ensure that the features are conditionally independent given the label sequence. This freedom makes CRFs particularly well suited to redundant input, as shown in [2]. For phonological feature and phonetic models, much of the available input is indeed redundant, providing further motivation for the approach. CRFs use the following exponential function to calculate the probabilities of various label hypotheses according to weighted feature functions:

$$P(\mathbf{y}|\mathbf{x}) \propto \exp \sum_{\mathrm{i}} (\mathrm{S}(\mathbf{x}, \mathbf{y}, \mathrm{i}) + \mathrm{T}(\mathbf{x}, \mathbf{y}, \mathrm{i})) \quad (3)$$

where $P(\mathbf{y}|\mathbf{x})$ is the probability of label sequence y given an input frame sequence x, i is the frame index, and S and T are a set of state feature functions and a set of transition feature functions, defined as:

$$S(x, y, i) = \sum_j \lambda_j s_j(y, x, i), \quad \text{and} \quad (4)$$

$$T(x, y, i) = \sum_k \mu_k t_k(y_{i-1}, y_i, x, i) \quad (5)$$

where λ and μ are weights determined by the learning algorithm. In natural language processing applications, the component feature functions s_j and t_k are often realized as binary indicator functions indicating the presence or absence of a feature, but in ASR applications we use real-valued functions, such as those derived from the sufficient statistics of Gaussians (e.g., [8]).

A more detailed description of this CRF paradigm can be found in [1], which shows that the results of phone recognition using CRFs is comparable to that of HMMs or Tandem systems, with fewer constraints imposed on the model.

In the past, we have trained CRFs using state features based on multi-layer perceptron posterior estimates of either phones, phonological features or both. In the current study, we additionally transform the multi-layer perceptron (MLP) posteriors with SVD to create further state feature functions. The properties of these state feature functions are discussed in Section 2.

2. DATA AND EXPERIMENTS

2.1. Acoustic Data and Neural Network Procedure

We use the TIMIT Speech Database [9] for our experiments. This resource is composed of read speech recorded in optimal conditions by 630 speakers. The sentences are designed to cover the widest possible range of phonetic sequences. Sentences are 3 seconds in length on average. The acoustic data is processed using a 25ms window and 10ms timestep. The corpus is hand-labeled at the phonetic level with 61 phonemes, including several labels for silence. The phone labels are propagated to each frame. We use the output of a 3-state HMM to determine the 3 state boundaries of each phone (except silence). This results in a phonetic state label set with 145 labels. We use the standard training, development, and core test sets, which have 3696, 400, and 192 sentences, respectively. We also test on an enhanced test set, which includes the core test set and an additional 752 unseen sentences. Results are shown for both the core and enhanced test sets.

Following the work in [1], we begin by extracting useful information from the acoustic input: we train MLPs to associate phone classes and phonological features with the acoustic data. We train a single MLP to discriminate between 61 phones (producing a posterior estimate of phone given acoustics), as well as a set of 8 other neural networks, each representing one phonological feature class. The output nodes of these 8 MLPs represent the specific features that each phonological feature class may express, and their values are posteriors. The 8 classes, which are sonority, voicing, manner, place, height, frontness, roundness, and tenseness, together express 44 features. The 44 phonological feature outputs and the 61 phonetic outputs are two redundant components of a single vector used to describe each data frame. The MLPs are trained on frame-level input described by 13 PLP features and double delta coefficients. With a context window of 9 frames, for each net there are 351 input nodes, 1000 hidden nodes, and a varying number of output nodes. We concatenate the output of the MLPs (105 features total) into a single vector per frame. The effectiveness of these feature-based posterior probabilities as state feature functions in a CRF framework was shown in [1] and [2]. Using these 105 features as input to the CRF system is our baseline (experiment 1). In experiments 2-8, these data form a matrix that undergoes SVD.

2.2. Latent Phonetic Analysis

We calculate the average posterior probability of each feature for each of the phone states over the training set. This results in a 105×145 matrix, with rows representing features, columns representing phone states, and the values representing the average MLP activation of that feature for that phone state. This matrix is far more dense than is typically seen in the document modeling paradigm. Nevertheless, SVD can still produce latent variables that better explain this data. We perform SVD on this matrix, resulting in three component matrices: $U_{105\times105}$ (features by dimensions), $S_{105\times145}$ (the weights attributed to each dimension), and $V^T_{145\times145}$ (dimensions by phones). We reduce the dimensionality to 50, the optimal value determined by tuning on the development set.

The left and right component matrices (U and V^T) are unitary (i.e $UU^T = I$ and $V^TV = I$) ; thus, $A = USV^T$ is equivalent to $S^{-1}U^TA = V^T$. Once we have constructed the inverse of the diagonal matrix S and the transposed feature matrix U^T, each frame of the original MLP data (A') is multiplied by $S^{-1}U^T$, thereby putting it in the "feature space". We calculate the cosine of the resulting feature space vector with each of the columns in the V^T matrix. This process results in 145 cosines for each frame of data, where the comparison of cosines indicates the relative affinity between that frame and each phone. We build the original matrices and perform SVD using training data. For training the CRFs, the frames of training data comprise A'. This means the 105 feature MLP data are projected to 50 features. At test time, the same matrices are used, but the test data make up A'.

We use the 145-element cosine vectors as input to a CRF system, which is trained using stochastic gradient descent to discriminate between a 48-label monophone set traditionally used in TIMIT phone recognition. Decoding is performed by first transforming the test data into cosine vectors, and then finding the most likely label sequence in the CRF using the Viterbi algorithm. The decoded labels are further collapsed to a standard 39 labels. This is experiment 2 in Table 1.

In experiment 3, we use the projected features directly as input to the CRF, rather than taking the extra step of calculating cosines. We also perform variance normalization on the projected features in order to scale data so that they lie mostly in the region $[-1, 1]$. The CRF training algorithm tacitly assumes that the data is normalized, because it uses the same learning rate for all features.

In experiment 4, we used the projected features to train a new neural network. The 50 projected features for each frame are mean and variance normalized, and presented to an MLP that discriminates between 48 phones. We also train CRFs against the combined outputs of several or all of these preprocessing steps (experiments 5 through 8).

Figure 1 shows a two dimensional embedding of the phones into the most significant SVD dimensions.[1] Here, we see that

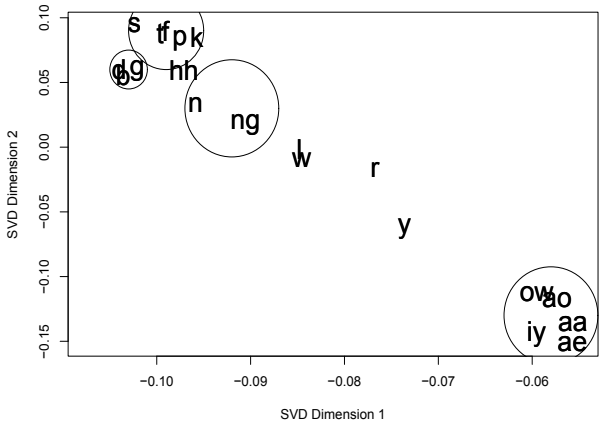

Fig. 1. Means of center states of 20 phones, projected to two dimensions via latent phonetic analysis

the two most important dimensions of Latent Phonetic Analysis distinguish between vowels and consonants, continuants and stops, and other groupings of consonants. In LSA, documents with similar topics group together in the same way ([4]). Similarly, grapheme strings with similar pronunciations are grouped together using an adaptation of LSA, ([10]).

3. RESULTS AND ANALYSIS

Experiment 2 in Table 1 indicates that the initial idea of applying the entire array of IR techniques, including taking the cosine to each phone-state "document," is not entirely successful. While this encapsulates most of the information contained in the baseline, it does not perform as well, and combining with the baseline provides no substantial improvement.[2]

Presenting the 50 reduced features to the CRF, alone or in combination with the baseline features (experiments 3 and 5), does not improve results over the baseline. The reduced space, which is a linear transformation of the input, is evidently too complex for the CRF to handle appropriately.

The results of experiment 4, with an accuracy on the enhanced test set of 72.0%, significantly improve upon the baseline results of 71.1%. Thus, we see a benefit from the application of the second non-linear transformation provided by applying the MLP to the output of LPA. The MLP trained on the LPA output further organizes the data in a way consistent with the labels we apply, by introducing the modeling of non-linear inter-relationships between different phone states.

When we combine the MLP features with the baseline features, as in experiment 6, we find improvements (72.6% accuracy) over the baseline and experiment 4 (the difference between 72.0% and 72.6% is significant at $p \leq .05$). This means that the features derived through the LPA analysis include information that was not readily present in the baseline

[1] A subset of phones are shown because the overlapping pattern prevents the full plot from being clear. The rest of the phones follow this pattern, with vowels on the lower right, and stops, nasals, etc. grouping together.

[2] A pure HMM baseline, as in [1], gets 67.32 accuracy.

Experiment Name	Input Size to CRF	Core Test Set		Enhanced Test Set	
		Acc.†	Corr.‡	Acc.	Corr.
1. Baseline	105	69.9	73.3	71.1	74.7
2. Project to 50 dimensions, take cosines	50	69.0	70.1	70.4	72.6
3. Project to 50 dimensions	50	69.5	71.8	70.8	73.3
4. Projected features to MLP	48	71.3	73.2	72.0*	74.3
5. Combine baseline w\ 50 dims	155	69.6	73.8	70.8	75.4*
6. Combine baseline w\ MLP output	153	71.8*	75.5*	72.6*	76.6*
7. Combine 50 dims w\ MLP output	98	71.7*	74.6*	72.6*	75.8*
8. Baseline, 50 dims, & MLP output	203	71.6*	75.9*	72.3*	77.1*
†=Accuracy ‡=Correctness		* = Significant at $p \leq .05$			

Table 1. Phone Recognition Accuracy and Correctness for all experiments

experiment. The new features provide useful information to the CRF both on their own and when combined with features that directly correspond to phones and phonological features.

In an additional experiment, we calculated the average activation of each feature over all frequent triphones, producing a 105x2598 matrix, and reducing it to 50 and 100 dimensions. We hypothesized that the LPA features would capture more information given a less compact matrix. We used the reduced normalized output as input to the CRF, with results slightly better than those in experiment 3. Using the reduced output to train an MLP produced worse results than those in experiment 4. When we combined the new features with the baseline, the results were worse than those in experiments 5-8. This was an unexpected result, but there are other ways of using SVD, including omission of the averaging across tokens. Some of these may contribute exploitable information.

Other dimensionality reduction techniques may achieve similar results; the Karhunen-Loeve transform, for example, is related to SVD (but assumes zero-mean data). We explored the use of the K-L transform as an alternative to SVD, finding equivocal results. While KLT applied to the means of the data is slightly worse than the baseline (68.8% accuracy on the enhanced set), feeding KLT features into the MLP corrects this deficiency, matching the LPA-MLP performance. While our use of SVD does not yield significantly greater results than other dimensionality reduction techniques, its success does spur us to explore related IR methods: the use of higher dimension matrix decomposition, used for cross-lingual IR [11], has shown encouraging results in this domain. A separate point to keep in mind is that traditional dimensionality reduction techniques in this domain operate over all frames of speech, whereas we average over phone states to create the SVD matrix, allowing us to conceive of groups of phones as documents, which seems to work better than considering all of the data. This is an area for more study.

4. CONCLUSIONS

We have found that latent phonetic variables, useful in distinguishing between phones, can be derived from acoustic data using singular value decomposition. By projecting the acoustic features into a reduced space, we can improve the results of our machine learning efforts. Furthermore, we have seen that the information derived from using this IR-inspired procedure is in some way complementary to the information derived from the use of MLPs over acoustic data. When the CRF takes into account both data sets, results improve over using either data set alone.

We will continue to explore both why and how adding features derived in this general fashion can improve phone recognition results.

5. REFERENCES

[1] J. Morris and E. Fosler-Lussier, "Further experiments with detector-based conditional random fields in phonetic recognition," in *Int'l Conf. on Acoustics, Speech, & Signal Processing (ICASSP-2007)*, Honolulu, Hawaii, 2007.

[2] I. Bromberg, J. Morris, and E. Fosler-Lussier, "Joint versus independent phonological feature models within CRF phone recognition," in *HLT-NAACL 2007*, Rochester, NY, 2007.

[3] T. K. Landauer, P. W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, vol. 25, pp. 259–284, 1998.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by Latent Semantic Analysis," *J'l of the Am. Society for Info. Science*, vol. 41, pp. 391–407, 1990.

[5] J. R. Bellegarda, "Latent semantic mapping," *IEEE Signal Processing Magazine*, vol. 22, pp. 70–80, 2005.

[6] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT, 1999.

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. of the 18th Int'l Conf. on Machine Learning*, 2001.

[8] A. Gunawardana, M. Mahajan, A. Acero, and J. Platt, "Hidden conditional random fields for phone classification," in *Interspeech*, 2005.

[9] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus," CDROM, 1993.

[10] J.R. Bellegarda, "Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy," *Speech Communication*, vol. 46, pp. 140–152, 2005.

[11] P.A. Chew, B.W. Bader, T.G. Kolda, and A. Abdelali, "Cross-language information retrieval using PARAFAC2," in *Proc. 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, 2007.