Articulatory Feature Recognition Using Dynamic Bayesian Networks

Joe Frankel, Mirjam Wester, Simon King

Centre for Speech Technology Research
University of Edinburgh
{joe, mwester, simonk}@cstr.ed.ac.uk

Abstract

This paper describes the use of dynamic Bayesian networks for the task of articulatory feature recognition. We show that by modeling the dependencies between a set of 6 multi-leveled articulatory features, recognition accuracy is increased over an equivalent system in which features are considered independent. Results are compared to those found using artificial neural networks on an identical task.

1. Introduction

In most ASR systems, the acoustic signal is described in terms of phones; words are simply concatenations of phone models. The notion that a word is a sequence of phone segments, i.e. the "beads-on-a-string" paradigm [1], makes it extremely difficult to model the variation that is present in spontaneous, conversational speech. Conventional systems use context-dependent phone models to deal with this variation. Articulatory features (AF) give an explicit representation of the asynchronous, overlapping nature of speech production, and therefore provide a means of deriving a principled model of the contextual variation characteristic of natural speech.

The number of studies that describe incorporation of articulatory features (also referred to as phonological features or articulatory-acoustic features) into ASR systems has been steadily increasing over the years. The approaches are quite diverse in terms of the type of data and models that have been used. The articulatory features can be derived from measured articulation [2, 3, 4], or generated from existing labels according to linguistic knowledge [5, 6, 7]. The models studied include artificial neural networks (ANN) [5, 6, 7], hidden Markov models (HMM)[5], linear dynamic models (LDM) [4], and dynamic Bayesian networks (DBN) [3, 8].

This work uses a set of multi-valued features, described in Section 2. Previous studies have shown some of the benefits of such features: reliable recovery from acoustic parameters using ANNs [6], noise-robustness [5], and less language-specific than phones [7]. A shortcoming of most previous approaches to AF recognition has been that features are modelled as statistically independent. This is an invalid assumption, and [7, 9] showed that place-of-articulation classification could be improved by training manner-specific models. Furthermore, when integrating feature recognition into ASR systems, the fact that accurate recognition of a given feature may be more important at some times than others has not been exploited. We propose that dynamic Bayesian networks provide an ideal framework within which to address these issues.

1.1. Dynamic Bayesian networks

A Bayesian network (BN) provides a means of encoding the dependencies between a set of random variables (RV). The RVs and dependencies are represented as the nodes and edges of a directed acyclic graph. A Bayesian network exploits missing edges (implying conditional independence) to factor the joint distribution of all random variables into a set of simpler probability distributions.

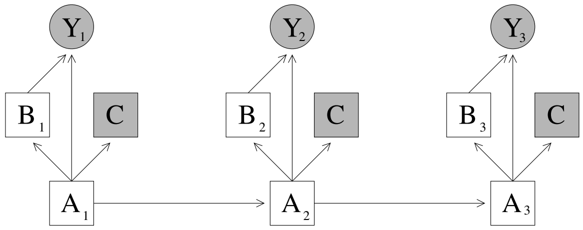

Figure 1: *Example Bayesian network showing dependencies between discrete random variables A, B, C and continuous observations Y. Round/square nodes show continuous/discrete variables, and shaded/unshaded distinguishes observed/hidden variables.*

A dynamic Bayesian network (DBN) consists of instances of a Bayesian network repeated over time, with dependencies across time. Such a model is shown in Figure 1 where each instance of the network is linked by conditioning A on its value at the previous time. The conditional independence structure represented by the model shows immediately that the joint probability of all variables at time 1 can be factored as:

$$p(A_1, B_1, C_1, Y_1) = p(Y_1|A_1, B_1)P(B_1|A_1)P(C_1|A_1)P(A_1) \qquad (1)$$

where $P()$ and $p()$ denote probability mass and density.

Dynamic Bayesian networks form a large class of models of which the HMM is one restricted case. These models provide an ideal framework to combine information from multiple sources, and offer the potential to build richer models of the parameterized speech signal than HMMs. One approach is the class of hidden feature models described by [8], in which articulatory features were used for feature-based factorization of the observation space. When the feature-based observation model was combined with a phone-based observation model, performance was improved on a simple recognition task. Used alone, the feature-based model did not outperform the baseline, which may be attributable to the simplifying assumption that features are independent given the phone state.

The ultimate goal of this work is a phone/word recognizer built around an articulatory feature factoring of the state and observation processes. Given such a broad class of models and so many design choices, we choose to split the task into a number of components. In the current work, we focus on feature recognition and tackle the issue of dependencies between different feature groups. In a further study, we intend to model the observation process in more detail, before going on to make the transition from frame-level feature accuracies to phone/word recognition.

2. Data

Experimental work uses the TIMIT corpus [10] following the standard train/test division, omitting the `sa` sentences (same for each speaker). Validation sets for the ANN and DBN experiments are described in [11] and [4] respectively, and the full test set (1344 utterances) is used for final evaluation. The acoustic waveform was parameterized as 12 Mel-frequency cepstral coefficients (MFCC) and energy, calculated every 10ms within 25ms windows. First and second order derivatives were appended giving 39 dimensional features which were used as acoustic input in all experimentation.

feature	values	cardinality
manner	approximant, fricative, nasal, stop, vowel, silence	6
place	labial, labiodental, dental, alveolar, velar, glottal, high, mid, low, silence	10
voicing	voiced, voiceless, silence	3
rounding	rounded, unrounded, nil, silence	4
front-back	front, back, nil, silence	4
static	static, dynamic, silence	3

Table 1: *Specification of the multi-leveled articulatory features. The right-hand column gives the cardinality of each feature.*

A set of 6 multi-leveled features as shown in Table 1 is used in this work. The feature groups are self-explanatory other than static, which gives an indication of rate of spectral change, such as occurs during, for example, diphthongs. Frame-level feature labels were generated from the TIMIT phone labels using mappings based on [12], and are similar to those described in [9].

3. Experiments

Artificial neural nets (ANNs) have been shown capable of recognizing articulatory features with high accuracy [5, 6]. To allow a direct comparison between ANNs and the DBNs with which this work is concerned, both ANN and DBN articulatory feature recognizers were trained on the same acoustic data and feature set. The following sections describe each of these in turn.

3.1. ANN feature recognition

A set of artificial neural networks was trained, one for each feature group, using the NICO Toolkit [13]. All networks are recurrent time-delay neural networks, consisting of three layers: an input layer, a single hidden layer, and an output layer. All networks used 100 hidden units, other than those for manner and place which used 200 and 300 respectively. During training, input-output pairs consist of frames of TIMIT acoustic features mapping to articulatory feature values. During testing, each network outputs an estimated feature value for a given acoustic frame. These can be interpreted as posterior probabilities, and in evaluating the network performance, the feature value with the highest associated probability is chosen.

Individual feature accuracies range from 78.3% for place, up to 92.9% for voicing, and the accuracy averaged over all features is 85.7%. The percentage of frames for which all features are correct together is 60.0%. These results are included for comparison with DBN performance and are repeated below in Table 5.

3.2. DBN feature recognition

The approach taken in this work is to commence with a baseline model in which the 6 feature groups are modeled as independent, add edges one at a time and evaluate each model on validation data. The model topology giving the highest accuracy will be chosen for final evaluation on the test set. During training, both acoustic and articulatory features are observed, and recognition uses a Viterbi search to find the most likely sequence of feature values given acoustic input. With a node representing each of the 6 feature groups, and given that DBNs are directed acyclic graphs, there can be at most $5+4+3+2+1 = 15$ within-frame edges. We compare two methods of choosing which dependencies to add, the first using information theoretic measures, and the second manually chosen.

3.2.1. *Observation model*

As discussed in [8], the requirement of specifying a distribution for every possible feature combination leads to problems of data sparsity. We follow the approach of [8] and adopt a factored observation model. For each of the six feature groups $F_1, \ldots, F_6$, a Gaussian mixture model is trained for each level $f \in F_k$. Using f_k to denote the level of feature F_k, the probability of an observation $\mathbf{y}$ is given as the product of the probabilities of $\mathbf{y}$ given the individual features:

$$p(\mathbf{y}|f_1, \ldots, f_6) = \prod_{k=1}^{6} p(\mathbf{y}|f_k) \qquad (2)$$

An observation model of this form will be used in all experiments, so that increases in accuracy can be attributable to adding dependencies between features. The total number of Gaussian distributions required is therefore 30, the sum of the cardinalities of individual features. Gaussian components were split and vanished during training using an adapted version of the scheme outlined in [14], and all covariance matrices are diagonal.

3.2.2. *HMM baseline and constrained silence models*

The baseline system, in which features are modeled as statistically independent is shown pictorially in Figure 2. Given the factored observation model, this amounts to 6 independent HMMs operating in parallel. With no constraints, such a system is capable of producing

dependencies	average correct	all correct together
independent (HMM)	80.1%	45.2%
independent, sync sil	81.1%	50.8%

Table 2: *Feature recognition validation accuracies for the baseline model in which features are independent, and also where all features are forced to synchronize recognition of silence.*

any of the $6 \times 10 \times 3 \times 4 \times 4 \times 3 = 8640$ feature combinations. One simple addition to the model is to ensure that recognition of silence is synchronized between features. A hidden discrete silence/non-silence node is added, and all features are forced to take silence/non-silence levels according to its value. Feature recognition results on

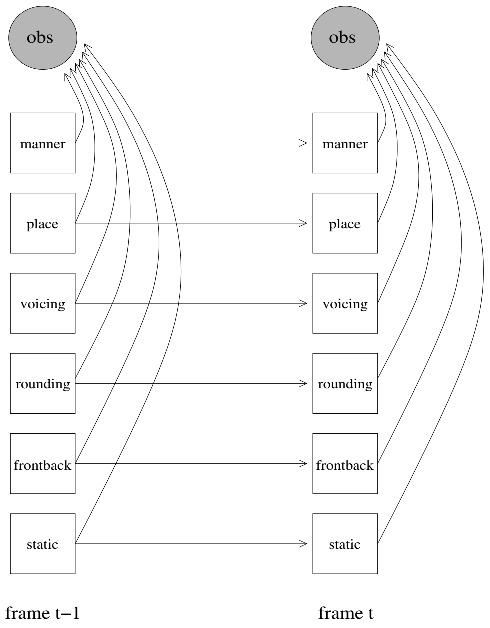

Figure 2: *Bayesian network depicting the baseline model in which features are modeled as independent.*

the validation set for the baseline and synchronized-silence models are shown in Table 2. Forcing common silence/non-silence decisions increases the average feature accuracy from 80.1% to 81.1%, and the percentage of frames where all features are correct simultaneously is increased from 45.2% to 50.8%.

3.2.3. Information-theoretic model selection

The mutual information $I(X;Y)$ between a pair of discrete random variables X and Y is calculated as:

$$I(X;Y) = \sum_x \sum_y P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} \tag{3}$$

and gives a measure of how much information one variable provides about the other. The conditional entropy $H(Y|X)$ is given by

$$H(Y|X) = -\sum_x \sum_y P(x,y) \log_2 P(y|x) \tag{4}$$

and indicates the uncertainty in Y given knowledge of X. Using the canonically-derived feature labels, the mutual information was calculated for each of the possible feature pairs, as were the conditional entropies for all feature combinations. The mutual information ranged from 0.43 for voicing and static up to 1.17 for manner and place, and the conditional entropy ranged from 0.13 for rounding given front-back up to 1.48 for place given static.

Table 3 shows validation accuracies for models where edges have been added one at a time in the order of ranked mutual information. The direction of the dependency follows that which gives the lowest conditional entropy. The dependencies which are added show that using the minimum conditional entropy favors conditioning a variable with lower cardinality on that with higher. We also tested the effect of choosing the direction of the edge such that the feature which gives the highest accuracy is the parent node, whilst adding dependencies in the order of ranked mutual information as

dependencies	average correct	all correct together
independent, sync sil	81.1%	50.8%
+ manner $\mid$ place	81.5%	52.4%
+ rounding $\mid$ front-back	81.7%	54.6%
+ front-back $\mid$ place	82.0%	56.2%
+ front-back $\mid$ manner	81.9%	56.1%
+ rounding $\mid$ place	81.9%	56.5%
+ rounding $\mid$ manner	82.1%	58.0%
+ static $\mid$ manner	82.1%	58.0%
+ voicing $\mid$ manner	82.1%	58.3%
+ voicing $\mid$ rounding	82.1%	58.3%
+ voicing $\mid$ front-back	82.1%	58.3%

Table 3: *Feature recognition validation accuracies: the order in which dependencies are added between feature pairs follows the ranked mutual information. The direction of dependency is chosen according to the minimum conditional entropy.*

in Table 3. The direction of the edges changes for all feature pairs other than for front-back and static conditioned on manner. Despite this, the results are very similar with an identical highest accuracy of 82.1% average and 58.3% all correct together. The direction of the edge in this case has minimal impact on recognition accuracy.

3.2.4. Manual model selection

The results reported in Table 4 are for models in which the order of adding dependencies is chosen according to the following: manner is considered the central parent node, and all other variables are conditioned upon it. Further dependencies are chosen in such a way as to maintain a balance in the size of the conditional probability tables, and hence free parameters, between variables. These results include the highest overall validation accuracy of 82.1% average and 58.5% all correct together. This model is shown pictorially in Figure 3 and is used for final evaluation on the test data.

dependencies	average correct	all correct together
independent, sync sil	81.1%	50.8%
+ place $\mid$ manner	81.5%	52.5%
+ front-back $\mid$ manner	81.8%	54.0%
+ rounding $\mid$ manner	81.9%	55.3%
+ voicing $\mid$ manner	82.0%	55.7%
+ static $\mid$ manner	82.1%	57.0%
+ rounding $\mid$ front-back	82.1%	57.8%
+ static $\mid$ place	82.1%	57.9%
+ front-back $\mid$ place	82.1%	58.1%
+ rounding $\mid$ place	**82.1%**	**58.5%**
+ front-back $\mid$ voicing	82.1%	58.5%

Table 4: *Feature recognition validation accuracies: the order in which dependencies are added is manually chosen. The overall highest validation accuracies are shown in bold face.*

Results for the test set are given in Table 5 and show that the DBN recognition performance is increased by introducing dependencies between features. Accuracies ranged from 71.9% for place

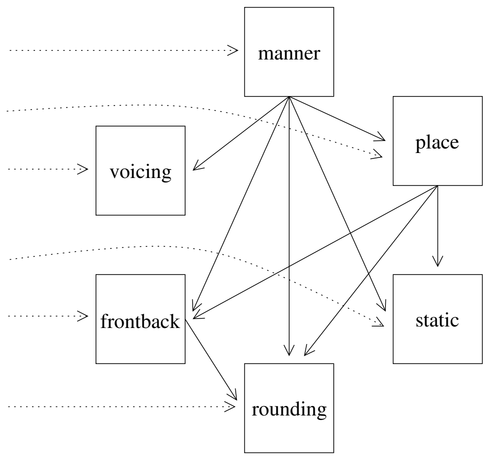

Figure 3: *Graph depicting the final model. Each feature is also conditioned on its value in the previous frame (implied by the dotted-line arrows) and a silence/non-silence node which, along with the observation process, has been omitted for clarity.*

to 89.4% for voicing, with an average of 81.5% and 57.8% of frames all correct together. Comparison between the ANN and DBN results shows that the ANN performs slightly better than the DBN. Nevertheless the results for the DBN are promising, especially given the simple observation model that we currently employ.

dependencies	average correct	all correct together
neural network	85.7%	60.0%
independent	80.8%	47.2%
independent, sync sil	80.6%	50.1%
dependent	81.5%	57.8%

Table 5: *Test set results.*

4. Discussion and future

Adding dependencies between variables provides a means of assigning a probability to each feature combination. This not only increases recognition accuracy, but decreases decoding time as unlikely combinations are assigned low or zero probabilities, and are therefore pruned or excluded from the search. The caveat which accompanies the results of Section 3.2, is that training on canonical labels leads to an overly strong set of constraints on feature co-occurrence. For the independent models, there were 3032 combinations found in the recognition output, reduced to 1084 where features were forced to synchronize use of the silence model. By contrast, there were only 50 combinations found in the output for the final model, suggesting that the introduction of so many constraints may reduce the system to a phone recognizer.

Canonically-derived feature labels provide a simple means of building an initial system, however the limitations here are clear. Future work will use embedded training in which the sequence of features is specified but not the timings of transitions. This approach will allow asynchronous feature changes, though in the absence of suitably detailed articulatory feature labels it is not clear how to evaluate such a system directly.

The observation model is overly simple, with features considered independent and a single Gaussian mixture model for each level. We intend to address this and develop an implementation using distributions specific to feature combinations where possible, backing off to product models where training data is limited.

Finally, we intend to use this topology as the basis for a phone recognizer with an underlying asynchronous feature layer.

5. Acknowledgments

Many thanks to Jeff Bilmes for such prompt and helpful responses to questions regarding GMTK, the toolkit with which the DBNs were implemented.

6. References

[1] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. of IEEE ASRU Workshop*, Keystone, CO., 1999, pp. 79–84.

[2] J. Sun, X. Jing, and L. Deng, "Data-driven model construction for continuous speech recognition using overlapping acoustic features," in *Proc. of ICSLP '00*, Beijing, 2000.

[3] T. Stephenson, H. Bourlard, S. Bengio, and A. Morris, "Automatic speech recognition using dynamic Bayesian networks with both acoustic and articulatory variables," IDIAP, Tech. Rep. 00-19, 2000.

[4] J. Frankel, "Linear dynamic models for automatic speech recognition," Ph.D. dissertation, The Centre for Speech Technology Research, Edinburgh University, 2003.

[5] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, 1999.

[6] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, pp. 333–353, 2000.

[7] M. Wester, S. Greenberg, and S. Chang, "A Dutch treatment of an elitist approach to articulatory-acoustic feature classification," in *Proc. of Eurospeech '01*, Aalborg, 2001, pp. 1729–1732.

[8] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic Bayesian networks," in *Proc. of Eurospeech '03*, Geneva, 2003, pp. 2529–2532.

[9] S. Chang, S. Greenberg, and M. Wester, "An elitist approach to articulatory-acoustic feature classification," in *Proc. of Eurospeech '01*, Aalborg, 2001, pp. 1725–1728.

[10] L. Lamel, R. Kassel, and S. Seneff, "Speech database development: Design and analysis of the acoustic-phonetic corpus," in *DARPA Speech Recognition Workshop*, 1986, pp. 100–109.

[11] M. Wester, "Syllable classification using articulatory-acoustic features," in *Proc. of Eurospeech '03*, Geneva, 2003, pp. –.

[12] P. Ladefoged, *A Course in Phonetics*, 2nd ed. Harcourt Brace Jovanovich, 1982.

[13] N. Ström, "Phoneme probability estimation with dynamic sparsely connected artificial neural networks," *The Free Speech Journal*, vol. Issue #5, 1997.

[14] J. Bilmes, *GMTK: The Graphical Models Toolkit*, SSLI Laboratory, University of Washington, October 2002.