EXPLICIT, N–BEST FORMANT FEATURES FOR VOWEL CLASSIFICATION

Philipp Schmid and Etienne Barnard

Center for Spoken Language Understanding
Oregon Graduate Institute, Portland, Oregon, USA
20000 N.W. Walker Road
Portland, OR 97291–1000, USA

ABSTRACT

We demonstrate the use of explicit formant features for vowel and semi–vowel classification. The formant trajectories are approximated by either three line segments or Legendre polynomials. Together with formant amplitude, formant bandwidth, pitch, and segment duration, these formant features form a compact feature representation which performs as well (71.8%) as a cepstral–based feature representation (71.6%). The combination of the formant and cepstral feature improves the accuracy further to 73.4%. Additionally, we outline future experiments using our robust, N–best formant tracker.

1. INTRODUCTION

Recent experiments on segment classification have focused on developing methods to represent the temporal and spatial correlations between cepstral features, such as mel–frequency cepstral coefficients (MFCC). Leung [1] and later Chigier et al. [2] used segment thirds and features designed to capture the changes across the segment boundaries to represent the dynamics of the feature trajectories. Osterndorf and Rokus [3] experimented with a mapping procedure for transforming the variable–length segments to a fixed–length feature representation. Later, Digalakis [4] used a dynamic system approach to model the temporal evolution of the MFCC features. Recently, Goldenthal introduced tracks [5], a non–parametric, fixed–length representation of the feature dynamics. All these approaches operate in the cepstral–feature domain, where the underlying articulatory dynamics are encoded in a highly non–trivial fashion. In contrast, we propose to use formant features as the feature representation of choice when attempting to model the dynamics of speech.

2. EXPLICIT FORMANT FEATURES FOR CLASSIFICATION

Formant features have a long history in speech recognition [6], but the accuracy and consistency of the formant tracking algorithms have generally been too low for high performance speech recognition. Talkin [7] lists some of the problems of formant tracking algorithms: (a) the determination of all formant candidates in each speech frame, (b) the enforcement of smoothness constraints across vowel/consonant boundaries (tight constraints within sonorants, loose constraints elsewhere in the absence of oral formants), and (c) the trade–off problem between trajectory smoothness and the amount of explained energy. In previous work, we have built a robust, N–best formant tracker [8], which addressed these problems. Firstly, the tracking algorithm uses a wild card mechanism which anticipates likely omissions of formant candidates and inserts wild cards to take their place in the tracking search. Secondly, formants are tracked only within sonorant regions (determined by a segmentation algorithm [9]), thus alleviating the problem of smoothness constraints across boundaries. Finally, instead of optimizing a global cost function which incorporates both smoothness and energy terms [10], our formant tracking algorithm maximizes a consistency score. Up to N consistent formant interpretations are passed on to the segment classification stage, thereby delaying a difficult decision (selection of the correct formant interpretation) until a later processing stage where additional information (e.g., classification probabilities, phonotactic constraints) is available. This basic idea of delaying difficult decisions as long as possible has proven very powerful in the context of frame–based search algorithms. By employing the same scheme for formant tracking

This work was supported through a grant in the Young Investigator Program of the Office of Naval Research. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

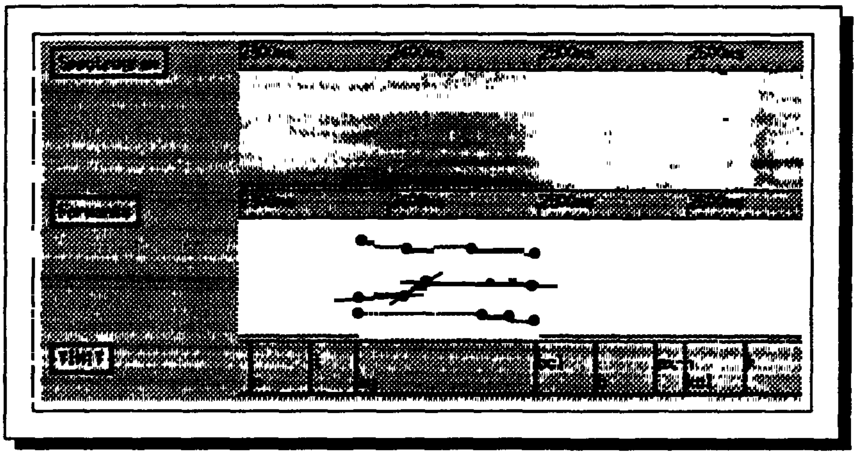

Figure 1: Formant Features: the second formant of the phoneme */ay/* is approximated by 3 line segments using the PLR algorithm. The dots indicate the locations of the formant features for this particular segment.

we hope that this approach will revitalize the use of formant–based features for classification.

3. FORMANT TRAJECTORY APPROXIMATION

Once the formant tracker has found a consistent formant interpretation, we are left with the problem of capturing the essence of the shape of each individual formant trajectory in order to extract features for the segment classifier. Various methods for cepstral or spectral feature spaces have been proposed in literature (e.g., [5, 3, 4]). In this work, we use two different methods for approximating the formant trajectories: line segments and Legendre polynomials.

Piecewise–Linear Regression

The Piecewise–Linear–Regression algorithm (PLR) [11] proposed by Krishnan and Rao is used to find the best three regression lines given a formant trajectory through an iterative method which converges generally after 3–5 iterations. The classification features are taken at the intersection points between adjacent line segments along with the start and end points. Additionally, the end points of the last (first) line of the adjacent segments (if they are sonorant) are also added to capture contextual information. An example is shown in Figure 1. The motivation for using line segments is the intention to extract direct feature measurements (e.g. formant locations) rather than presenting the classifier with an encoding of the salient information. They also lend themselves more conveniently to speaker normalization techniques.

Legendre Polynomials

The formant trajectories can also be approximated by polynomials. We chose third–order Legendre polynomials (LEG) because of their use of an orthonormal basis, which produces more robust estimates of the coefficients of the polynomials.

The third–order basis polynomials ϕ_j in Equation 1 are taken from [12]. The Legendre coefficients a_j are then computed as follows:

$$a_j = \frac{1}{M+1} \sum_{i=0}^{M} y_i \cdot \phi_j(\frac{i}{M}) \quad . \qquad (1)$$

An arbitrary point $x \in (0,1)$ within the scaled segment can be reconstructed by:

$$f(x) = \sum_{j=0}^{3} a_j \cdot \phi_j(x) \quad , \quad x \in (0,1) \quad .$$

The average prediction error can then be computed as:

$$PE = \frac{1}{M+1} \sum_{i=0}^{M} | \, y_i - f(\frac{1}{M}) \, | \quad .$$

As in the PLR case, contextual features are extracted if the neighboring segments are sonorant (and hence are also approximated by Legendre polynomials).

4. VOWEL/SEMI–VOWEL CLASSIFICATION EXPERIMENTS

In this section, we report on a series of classification experiments using the TIMIT database to determine the

Feature Set	# Features	Accuracy
PLR (3 Line Segments) + D	19	**66.0%**
PLR + D + A + P	23	**68.2%**
LEG + PE + D	22	**69.6%**
LEG + PE + D + A + P + B	29	**71.8%**
MFCC + D	113	**71.6%**
MFCC + D + P	114	**73.1%**
MFCC + LEG + PE + D + A + P + B	141	**73.4%**

Table 1: Vowel/Semi-Vowel Classification Results

potential of explicit formant features such as trajectory approximation schemes, formant amplitude, and formant bandwidth for vowel/semi-vowel classification. Additionally, we investigate the use of pitch estimates, since pitch information is often used in proposed speaker normalization methods motivated by perceptual experiments [13].

The classification experiments use the 14 vowels and 4 semi-vowels (liquids and glides) of the final phoneme set as defined by Lee in his initial TIMIT classification work [14]. The classifier, a multilayer perceptron (MLP) trained with conjugent-gradient optimization, was trained on the NIST training set containing 3698 utterances (si,sx) for a total of 58268 training tokens. The test set consisted of 400 utterances from the TIMIT test corpus (6462 test tokens). In addition to the PLR and LEG features, we also added formant amplitude (A), formant bandwidth (B), pitch [15] (P), the prediction error (PE) in the case of the Legendre polynomials, and the segment duration (D) to the feature set. These initial classifiers were trained with $N = 1$. The best feature configuration achieved 71.8% classification accuracy, as can be seen from the summary of the classification results in Table 1.

For comparison, we trained an MLP classifier using MFCC features: (a) MFCC averaged over segment thirds, (b) MFCC averaged over the 2 frames left (right) of the left (right) boundary, (c) MFCC averaged over 3 frames starting 2 frames to the left (right) of the left (right) boundary, as well as the segment duration. As can be seen from Table 1, the best feature configuration based on formant features performs almost identically to the MFCC-based classifier (71.6%).

The best classification performance of 73.4% is achieved when both feature sets are combined. However, adding only pitch (P) to the set of MFCC features increased the classification accuracy to 73.1%, indicating that pitch was responsible for at least the major part of the improvement in performance.

5. N-BEST EXPERIMENTS

The above results were obtained using $N = 1$. However, our formant tracking algorithm produces a list of up to N consistent interpretations of the formant information for each segment or sequence of sonorant segments. In the N-best classification paradigm, the phonetic category a^* with the highest probability over all N interpretations $Interp_k$ is used to label the segment:

$$a^* = \operatorname*{argmax}_{a,k} p(a|Interp_k) \quad \text{for k} = 1 \ldots \text{N} \qquad (2)$$

That is, we address the question: "Assuming this is the correct formant interpretation, which phoneme would it be?" Note that we propose to use more than one set of features per segment in the classification process! This is a novel concept, as far as we know.

To avoid time-consuming labeling of formants, we use an iterative method for labeling (marking the correct interpretations) the data and training of a vowel classifier at the same time. Firstly, an initial classifier is trained for $N = 1$. Secondly, this initial classifier is then used to "label" the formants by selecting the interpretation that yields the highest probability for the correct vowel category. A new vowel classifier is then trained on the machine-labeled interpretations. The labeling and training steps are repeated until the performance of the classifier on a separate development test set converges.

Table 2 summarizes the results of the iterative training process using PLR features. The training set consisted of half of the TIMIT training utterances. The remaining training utterances were evenly distributed

Itr	1	2	3	4	5	6
Acc	62.7	63.6	63.8	63.8	64.3	63.6

Table 2: Classification Results for Iterative Training Procedure

into two development test sets. The number of formant interpretations per segment was set to 3. The iterative process reduced the error rate by 2.5%. This small decrease is mainly due to the fact that most of the interpretations are actually identical, therefore reducing the choice for the selection of training patterns. Improving the pruning strategy of the formant tracker should result in a higher performance gain for the iterative training procedure.

6. CONCLUSIONS AND FUTURE WORK

This research shows that formant-based features can produce classifiers that are as accurate as cepstral-based classification features with a 4-fold reduction in the size of the feature space. We now intend to integrate the N-best tracker into our phonemic segment classifier.

This work opens up the possibility of incorporating perceptually motivated dynamic features (which are almost invariably described in terms of formant motions) into automatic classifiers. Future work will concentrate on investigating explicit normalization methods as well as such dynamic phenomena, as suggested in the literature on vowel perception. Additionally, consonant classifiers using formant-based features will be built.

7. REFERENCES

[1] Hong C. Leung. *The Use of Artificial Neural Networks for Phonetic Recognition*. PhD thesis, Dept. of Electrical Engineering, MIT, 1989.

[2] B. Chigier, and H. Leung. The Effects of Signal Representations, Phonetic Classification Techniques, and the Telephone Network. In *Proceedings of ICSLP*, pages 97–100, Banff, Canada, 1992.

[3] Mari Ostendorf and Salim Roukos. A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12), December 1989.

[4] Vassilios V. Digalakis. *Segment-based Stochastic Models of Spectral Dynamics for Continuous Speech Recognition*. PhD thesis, Boston University, Boston, MA, 1992.

[5] W. Goldenthal. *Statistical Trajectory Models for Phonetic Recognition*. PhD thesis, Department of Aeronautics and Astronautics, MIT, 1994.

[6] D. Klatt. Review of arpa speech understanding project. *Journal of Acoustical Society of America*, 62:1345–1366, 1977.

[7] David Talkin. Speech Formant Trajectory Estimation Using Dynamic Programming with Modulated Transition Costs. AT&T Internal Memo MH 11222 2924 2D-410, AT&T, 1987.

[8] P. Schmid, and E. Barnard. Robust, N-best Formant Tracking. In *Proceedings of 4th European Conference on Speech Communication and Technology*, volume 1, pages 737–740, 1995.

[9] P. Schmid. *Explicit, N-Best Formant Features for Segment-Based Speech Recognition*. PhD thesis, Oregon Graduate Institute of Science and Technology, 1996.

[10] Y. Laprie. A New Paradigm for Reliable Automatic Formant Tracking. In *Proceedings of ICASSP*, pages 201–204, 1992.

[11] S. Krishnan and P. V. S. Rao. Segmental Phoneme Recognition using Piecewise Linear Regression. In *Proceedings of ICASSP 1994*, pages I-49 – I-52, 1994.

[12] S. Chen and Y. Wang. Vector Quantization of Pitch Information in Mandrin Speech. *IEEE Transactions on Communications*, 38(9):1317–1320, 1990.

[13] J. Miller. Auditory-perceptual interpretation of the vowel. *J. Acoustical Society of America*, 85(5), 1989.

[14] K. F. Lee. *Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*. PhD thesis, Computer Science Department, Carnegie Mellon University, 1988.

[15] S. van Vuuren. Pitch detection. Technical report, Dept. of Electrical and Electronic Engineering, University of Pretoria, South Africa, 1992.