SMOOTHNESS ANALYSIS FOR TRAJECTORY FEATURES

Zhihong Hu and Etienne Barnard

Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, 20000 N.W. Walker Road, P.O. Box 91000, Portland, OR 97291-1000, USA, (zhihong@cse.ogi.edu)

ABSTRACT

Dynamic modeling of speech is potentially a major improvement on Hidden Markov Models (HMMs). In one approach, trajectory models[1] are used to model the dynamics of the spectrum, and are used as basis for classification [1, 2]. Although some improvement has been achieved in this way, one would hope for more substantial improvements given that the independence assumption is removed. One reason why this was not achieved may be that the trajectory models are based on cepstral coefficients; we show that these tracks contain spurious oscillations. This suggests that these trajectory features might have a high within-class variance. We introduce a measure of evaluating the smoothness of trajectory-based features. This measure provides a method of selecting the best of a set of similar features. Formant trajectories prove to be significantly smoother than trajectories of mel scale cepstral coefficients (MFCC) by this measure, but this does not translate directly to improved performance.

1. INTRODUCTION

It is well known that speech has high variability. However, consistency exists both within a phonetic unit and across an utterance. In general, in a given phonetic context, each phone has a certain configuration of articulators associated with it. Although there are variances between different speakers, they share the same movement of the articulators when producing the same phone sequence. The similarity in dynamic movements of the articulators corresponds to the similarity of dynamics in the acoustics of each phone. Therefore, features that can capture these dynamic movements in the acoustic signal should be very useful in speech recognition.

Several trajectory models[1, 3, 4, 5] have been proposed to capture the dynamic behavior of speech. In Goldenthal's work[1], for instance, the temporal behavior is modeled by templates of the dynamics of the acoustic attributes used to represent the signal. By estimating their spatial-temporal correlation structure, trajectory models are generated for phonetic recognition. Most of Goldenthal's results are obtained using trajectories of mel scale cepstral coefficients (MFCCs).

In our past efforts to incorporate this dynamic information in speech recognition[6], we built a recognition system using syllable-like units and used the statistical trajectory models (as described in [1]) as the main features. Although we obtained encouraging results, we found that trajectories of the coefficients in a cepstral-based analysis oscillate through time even when the signal is changing slowly and smoothly. This is caused by the trigonometric mapping between cepstral coefficients and frequency components (Section 3). This seems to indicate that it may be advantageous to use a description more directly related to the mechanisms of speech production in a trajectory-based model – e.g. a description based on formants.

However, it is not entirely straightforward to compare the trajectories produced by two such dissimilar feature sets. We have therefore designed a statistical measure to compare various trajectory-based feature representations.

In the following sections we describe the definition and the underlying meaning of the measurement (Section 2), the simulation of the MFCC oscillations (Section 3), and the measurement results we obtained on TIMIT vowels (Section 4). Finally, in Section 5, we summerize our work and propose related future work.

2. SMOOTHNESS MEASUREMENT

In order to compare the relative smoothness of various trajectories, we define a measure called relative percentage error (RPE):

$$\mathrm{RPE} = \frac{\sum_{i=1}^{n}(T(i) - P(i))^2}{\sum_{i=1}^{n}(T(i) - \mu)^2}$$

where T denotes the n state trajectory for the specific feature component. (That is, the trajectory is represented by n sample values, and the summation runs over these samples.) P denotes a low-order polynomial fit to this trajectory, and μ denotes the mean value of the trajectory. The numerator

$$\sum_{i=1}^{n}(T(i) - P(i))^2$$

reflects how well the low-order polynomial fits the measured track, and the denominator normalizes this fitting error by the overall signal variance.

This measure therefore tells us how well the low-order polynomial fit models the measured trajectories; we have experimented with second-order and third-order fits with almost identical results in both cases.

3. SIMULATION EXPERIMENTS

In this section we describe a simulation experiment which explains the observed oscillations in the MFCC trajectories graphically. The aim of this experiment is to observe the behavior of the MFCC trajectories when the formants move across multiple frequency bands.

A signal with two rising tones is generated by adding two sinusoidal signals. Trajectories of the MFCC coefficients are calculated for this signal. The spectrogram and the trajectories are presented in Figure 1. In Figure 1, the first window shows the spectrogram of the signal and the following six windows show the trajectories of the 0th, 1st, 2nd, 3rd, 4th, 5th and 6th MFCCs, respectively.

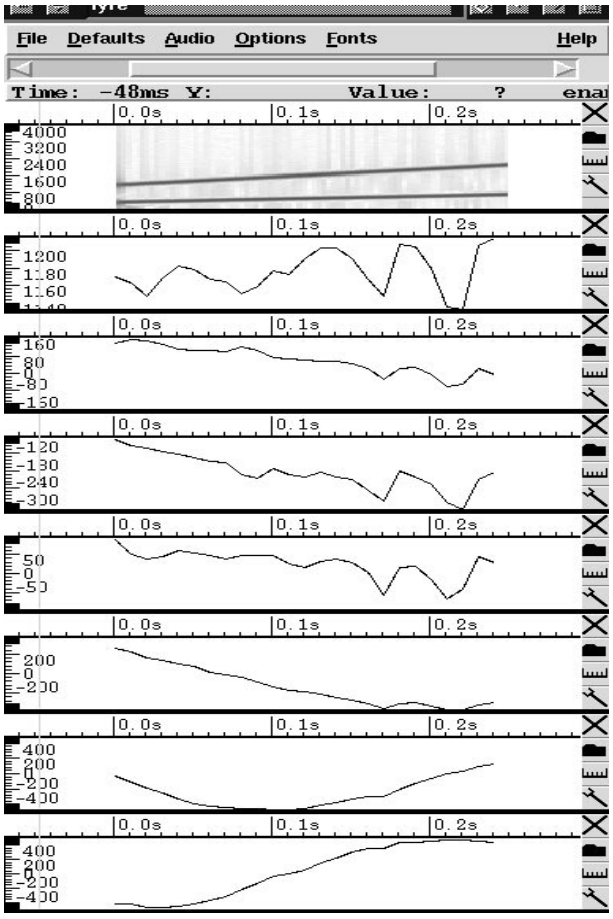

Figure 1. Spectrogram and the MFCC trajectories of signal with two rising tones.

We see that a signal as simple as a pair of rising tones can cause significant oscillation of the MFCC trajectories - especially the higher-order coefficients vary through a substantial fraction of their total range despite restricted changes to the signal. This fact is confirmed by analysis of the definition of MFCCs, and suggests that smoother trajectories may be obtained with a more "natural" feature set.

To further illustrate this observation, Figure 2 shows a real example from the TIMIT database. In Figure 2, the first window shows the spectrogram of the signal and the following six windows show the trajectories of MFCCs 0 through 6. Comparable variations are observed.

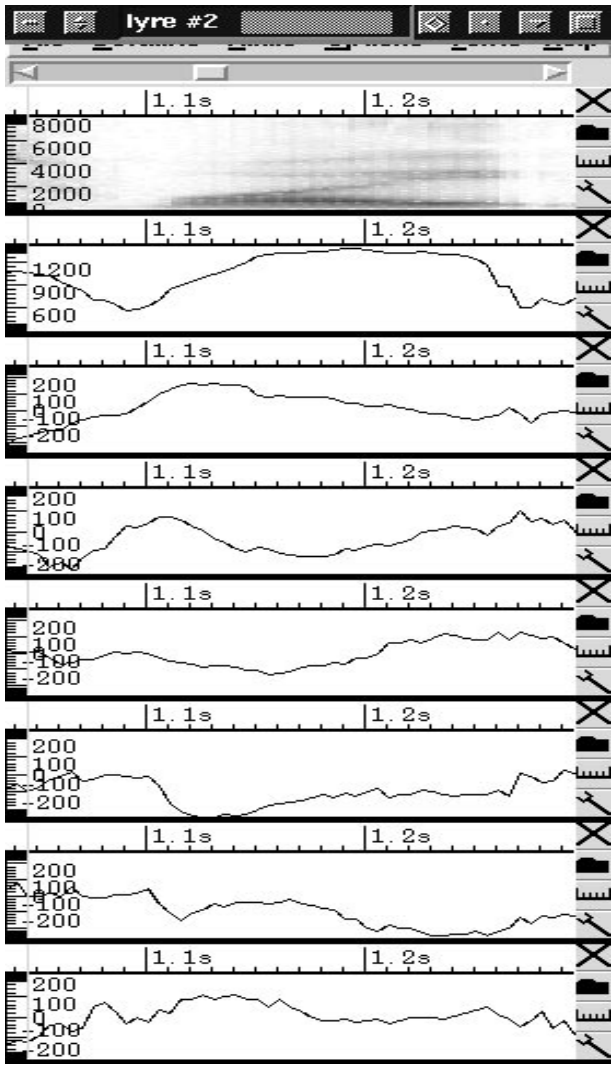

Figure 2. Spectrogram and the MFCC trajectories of signal for a TIMIT example (phoneme "ay").

4. SMOOTHNESS ANALYSIS AND EXPERIMENTS

This hypothesis has been tested with a set of experiments designed to compare the smoothness of formant trajectories and MFCC trajectories. Classification results using these trajectory features are also presented. In these experiments, the task is context-independent, gender-independent vowel classification. The TIMIT database is used as the corpus. The 16 vowels used in the experiments are:

`aa ae ah ao aw ay eh er ey ih iy ow oy uh uw ux`

The training set includes all the *sx* and *si* files in the TIMIT training set, and the test set includes all the *sx* files in the TIMIT test set. Details are shown in Table 1:

Data Set	#Utterances	# vowels
train	3696 `sx si`	31863
test	840 `sx`	6771

Table 1. The data set used in the experiments.

In the experiments performed, three formants are estimated for each vowel by using a formant estimation method proposed by Welling and Ney[7]. The results of the formants smoothed by a median filter of window width 5 are also presented. These are compared to 14th-order MFCCs.

4.1. Smoothness comparison

Tracks with 10 states for each dimension (MFCC or formant) are computed for each segment. RPE values are calculated for each track. Statistical tests (T test)[8] are then performed to verify the assumption that the RPE value of the formant trajectories are smaller than that of the MFCCs. The t value is calculated as:

$$\mathrm{t} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

where $\bar{X}$ denotes the mean of all the formant trajectories' RPE value, μ_0 denotes the mean of all the MFCC's RPE value, and S denotes the sample standard deviation of all the formant trajectories' RPE values; n is the number of samples. If we evaluate significance at the 0.5% level, using a one-tailed test, we need $|t| > 2.576$ for a significant difference.

The statistical smoothness comparison result is shown in Table 2. The numbers in the table are the t values for different phonemes.

phoneme	formant	smoothed formant
`aa`	−11.2068	-34.1291
`ae`	-12.4169	-31.4355
`ah`	-12.311	-32.8983
`ao`	-12.1021	-33.072
`aw`	-4.9549	-12.6598
`ay`	-11.200	-23.6606
`eh`	-19.2256	-49.579
`er`	-10.3508	-26.1233
`ey`	-16.2408	-37.9998
`ih`	-18.0738	-51.9455
`iy`	-18.073	-56.479
`ow`	-13.947	-29.7495
`oy`	-8.06058	-10.8648
`uh`	-4.14614	-14.1025
`uw`	-5.31656	-13.1767
`ux`	-6.0073	-22.3937

Table 2. Statistical smoothness comparison result using Relative Percentage Error measure.

Clearly, all of the RPE values for the formant features are significantly smaller than that of the MFCCs. This suggests that the MFCC feature set introduces significant variation which is modeled in a simpler fashion by formants. Consequently, it is reasonable to expect that within-class variances for the formant trajectory features are smaller than that of the MFCCs.

4.2. Classification using trajectory features

Neural networks were trained to be classifiers using the trajectory features for both MFCCs and formants. We chose to use 10 states for each trajectory in this experiment. The trajectory of the energy and the log duration of the segment are also used as part of the feature set. The classification results are shown in Table 3; improved performance is not obtained with the formants, despite their smoothness.

feature	MFCC	formant	smoothed formant
dimension	151	41	41
% correct	66.6%	64.6%	64.5%

Table 3. Classification results using different trajectory features.

4.3. Classification using polynomial approximations

The smoothness analysis has shown that the formants are a smoother representation. This suggests that coefficients of a polynomial approximation of the formant trajectory are suited to be used as features for classification.

Figure 3 shows the polynomial fitting of a typical formant trajectory. In the figure, the first window shows the spectrogram and the original formant position estimation. The second window is the phonetic label. `l-ay+ih` means phoneme `ay` in the left context of `l` and right context of `ih`. The third window shows the formant trajectory reconstructed from the 3rd-order polynomial fitting.

In this experiment, the coefficients of an orthonormal polynomial approximation (the Legendre polynomial[9]) for the trajectories are used as features in classification.

The classification results are shown in Table 4.

feature	MFCC	formant	smoothed formant
dimension	61	17	17
% correct	69.7%	66.7%	66.8%

Table 4. Classification results using polynomial approximation for different trajectories.

The results shows significant improvement over the trajectory features, for all cases. The significant improvement on MFCC features might be attributed to the dimension reduction in feature space which helps in the neural network training process.

Despite the smaller variance of the formant features, we did not obtain better performance on this classification task. This may be related to the additional information contained in the cepstral features (of which there are almost four times as many as the formant features). Note that our results are comparable to what Goldenthal reported in [1].

4.4. Classification using more information

To investigate whether additional information can be used to improved the performance of the formant-based models, features describing other attributes of the formants are added into the feature set to be investigated. The average bandwidth of the formants and contextual information are tested in these experiments. The contextual features are calculated as the average formant location of the three frames to the left of the left boundary and three frames to the right of the right boundary of the segment.

The classification results are shown in Table 5.

feature	+bandwidth	+context
dimension	20	28
% correct	67.2%	68.0%

Table 5. Classification results using feature represent other information.

These results show that adding relevant information can help improve classification performance. We found no improvement for classification performance for MFCC features after adding the contextual features. Thus, it seems as though the additional spectral information contained in MFCCs can indeed account for at least some of the performance differences observed.

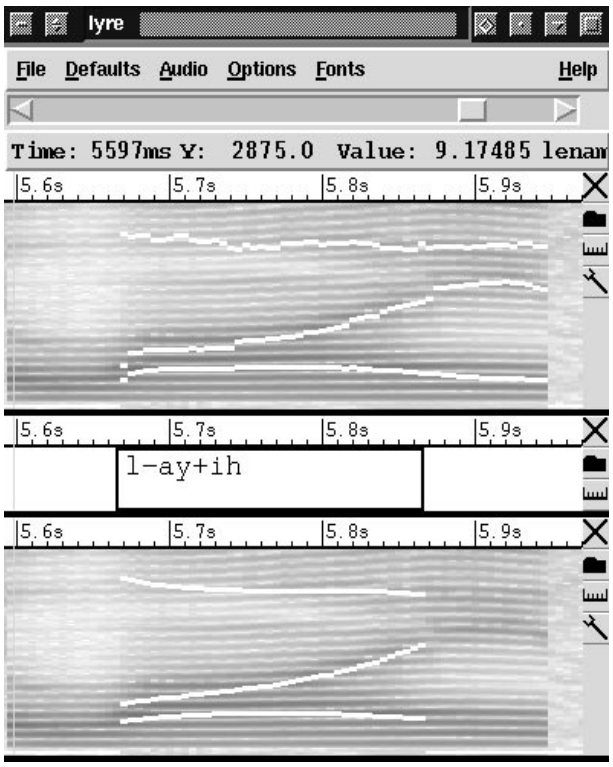

Figure 3. Polynomial fitting of the formant trajectory

5. SUMMARY AND FUTURE WORK

We presented a new measure for comparing the smoothness of different trajectory features. Preliminary experiments show that trajectories of formants are smoother than those of MFCC coefficients.

These results do not translate directly into improved performance at vowel classification. Adding additional information helps to improve the classification performance. Further research is needed to see whether other information can be used to obtain even further improvement. We also need to investigate whether the imperfections of the formant-tracking algorithm are responsible for a substantial degradation in performance.

To complete the study of the trajectory features, we will further investigate the effects of moving formants on the cepstral-based analysis technique and study possible integration of related features (such as trajectories of the formant bandwidth and other features describing the trajectory shape) to formant trajectory features.

Acknowledgement: This work was jointly supported through NSF/ARPA grant 107 and 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.

REFERENCES

[1] W. Goldenthal, *Statistical Trajectory Models for Phonetic Recognition*. PhD thesis, M.I.T., Auguest 1994.

[2] M. Afify, Y. Gong, and J. Haton, "Estimation of mixtures of stochastic dynamic trajectories: application to continuous speech recognition," *Computer Speech and Language*, no. 10, pp. 23–36, 1996.

[3] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition," *IEEE Transaction on Accoustics, Speech, and Signal Processing.*, vol. 37, no. 12, pp. 1857–1869, 1989.

[4] V. Digalakis, J. Rohlicek, and M. Ostendorf, "A dynamical system approach to continuous speech recognition," in *ICASSP-91*, pp. 289–292, 1991.

[5] V. Digalakis, J. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the em algorithm and its application to speech recognition," *IEEE Transaction of Speech and Audio Processing.*, vol. 1, no. 4, pp. 431–442, 1993.

[6] Z. Hu, S. Schalkwyk, E. Barnard, and R. Cole, "Speech recognition using syllable-like units," in *ICSLP-96*, October 1996.

[7] L. Welling and H. Ney, "A model for efficient formant estimation," in *ICASSP-96*, pp. 797–800, May 1996.

[8] R. Hogg and R. Tanis, *Probability and Statistical Inference*. Macmillan Publishing Company, 1993. General Intro : ISBN 0-02-355821-0.

[9] S. Chen and Y. Wang, "Vector quantization of pitch information in mandarin speech," *IEEE Transactions on Communications*, vol. 38, no. 9, pp. 1317–1320, 1990.