EFFICIENT ESTIMATION OF PERCEPTUAL FEATURES FOR SPEECH RECOGNITION

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

A number of studies have shown that a pair of perceptual effective formants can be defined to capture most of the phonetic information present in vowels. Various methods of computing the effective formant values were proposed. However, many of them depend on the accuracy of conventional formant estimation. In this work, we study methods of automatically estimating perceptual effective formants without estimating the actual formant values and compare the results with the perceptually measured effective formant values. The preliminary results show that the method is effective in estimating the perceptual effective formants. Classification experiments using perceptual effective formants as explicit features do not demonstrate any advantages. However, using the perceptual effective second formant value as input to our formant estimation algorithm can help to correct up to 44% of the formant tracking errors.

1. INTRODUCTION

Substantial improvements to speech recognition may be attained if the dynamic properties of spoken language are modeled adequately[1]. Several lines of reasoning indicate that formant-based features[2] are a good candidate for this goal, at least for the sonorant regions of speech. Using the formant frequencies themselves for recognition suffers, however, from a number of significant drawbacks. First, the formants are highly dependent on vocal tracks, therefore not speaker independent, and (more importantly for practical implementations) have been notoriously difficult to track reliably.

Fortunately, there is much evidence that complete knowledge of all formant frequencies is not required for accurate recognition. Perceptual experiments by Fant and others[3, 4] suggest that a two-formant approximation model (perceptual effective formants) is a valid and robust representation for most vowels. [1] In their work, two prominent spectral peaks (they use F1 and F2' which is a function of the first four formants.) are found to be sufficient to describe all Swedish vowels[5]. The detailed description of the model, as well as measurements of the effective formants of the 18 cardinal vowels, is provided in [6]. Fant and Risberg[3] studied vowel separation using the "effective" formant model , comparing it to the pure F1 and F2 formant values. These studies suggest that the effective formant model (F1-F2') separates the vowel space better than the standard F1-F2 combination. In addition, these models do not require detailed tracking of formant frequencies: when resonant frequencies become so close that it is hard to distinguish the individual tracks, a single effective formant is deemed sufficient for recognition. These previous studies suggest that perceptual effective formants might be good speaker independent features to describe vowels.

Based on these insights, in this paper, we investigate a method of estimating perceptual effective formants using low-order PLP spectral peaks on real speech and study the effectiveness of using them as features in vowel classification.

2. EFFECTIVE ESTIMATION OF PERCEPTUAL FORMANTS

Various methods of computing the F1' and F2' values have been proposed. Most commonly, these values are computed as functions of the actual formant values[3, 4, 7]. As noted, this requires determination of the underlying formant tracks, and thus is undesirable for practical systems.

Itahashi and Yokoyama[8] estimate the effective formants without computing the true formant frequencies. In this method the high-order LPC spectrum is warped according to the Mel scale. The resulting power spectrum is then weighted according to an equal loudness contour, from which the autocorrelation function is computed by the inverse Fourier transform. Linear predictive analysis is performed to obtain the poles which represent the formants according to the Mel scale. The second formant thus estimated was found to be a good estimate of F2' determined by auditory matching. Hermansky[9] performed similar experiments, studying the spectral peaks obtained using a 5th order PLP model. These experiments were performed using synthesized speech for each of the 18 cardinal vowels. The synthesized data were based on formant values provided by Bladon and Fant[6]. The agreement between values from the PLP model and from Bladon and Fant's perceptual data is rather good.

We intend to build on these insights, with the eventual goal of using effective formants as input to a speech-recognition system. With that goal in mind, we here evaluate various methods of calculating the effective formant frequencies.

For our comparison, we use the set of 18 vowels analyzed by Bladon and Fant[6]. (Whereas[9] used synthe-

[1] In this paper we refer the effective perceptual formants as as F1-prime (F1') and F2-prime (F2').

sized versions of these vowels, we were able to obtain the actual stimuli.) In the following table we compare estimated values for F2' obtained with various estimators to the values measured in Bladon and Fant's perceptual experiments. These estimators were based on (a) standard linear predictive (LP) analysis, (b) LP analysis based on mel-scale frequency warping (MEL),(c) Itahashi's method (MELP), and (d) LP analysis based on Bark scale frequency warping (PLP). In each case, a sixth-order analysis was performed[2], and the value for F2' was defined as the frequency corresponding to the second pole of the corresponding polynomial. Also included in the table are values obtained with the formant-based calculation proposed by Bladon and Fant (fmt). All the values in the table are in Bark scale.

phone	percept	fmt	LPC	MEL	MELP	PLP
i	14.1	14.2	14.5	13.6	14.0	12.1
e	12.5	12.5	12.4	11.6	11.8	11.4
E	11.7	11.3	10.1	10.5	10.4	11.0
a	9.7	9.5	8.5	8.8	8.9	9.4
A	8.2	8.6	8.0	7.8	8.2	8.9
O	6.6	6.8	14.6	6.6	7.3	8.7
o	6.0	6.2	14.4	8.0	4.6	8.1
u	5.8	6.0	15.1	8.2	9.2	6.9
y	11.8	11.9	11.8	10.9	12.1	11.4
0	10.1	10.2	8.4	9.5	9.5	10.1
oe	10.4	10.3	8.1	9.3	9.5	9.9
OE	9.7	9.8	8.1	9.1	9.4	9.7
@	7.4	8.1	7.8	7.9	8.0	8.8
^	9.0	9.0	14.8	8.5	8.2	9.2
g	9.2	9.3	16.4	8.6	8.7	9.6
m	9.1	9.3	8.5	8.9	8.9	9.7
I	10.8	10.8	10.3	10.3	10.1	10.9
U	9.9	9.8	9.8	10.0	9.9	10.5
avg err	-	0.18	2.7	0.76	0.73	0.79

Table 1. Perceptually estimated (Bladon and Fant, 1978) and automatically estimated frequencies of perceptual effective formants of 18 cardinal vowels in Bark scale.

We see that analysis methods which include spectral auditory scale warping are substantially more accurate than pure linear-predictive analysis, but not as accurate as the formant-based method. In particular, the PLP method overestimates F2' for /o/ and /O/, and underestimates it for /i/. The comparative rankings are confirmed by the values obtained for F1'. In addition, LPC sometimes fails to produce a pole below 1000 HZ.

3. APPLICATION OF PERCEPTUAL EFFECTIVE SECOND FORMANT

To explore the application of the perceptual effective formants, experiments have been performed to test various possibilities. In the following classification experiments, the task is context-independent, speaker-independent vowel classification. The TIMIT database is used as the corpus. The 14 vowels used in the experiments are:

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

[2] 5th order PLP model was used first and found that when the F1' and F2' merge, it is difficult to determine automatically.

The training set includes all the *sx* and *si* files in the TIMIT training set, and the dev set is MIT dev set[10] which contains *sx* and *si* files. Details are shown in Table 2:

data set	# utterances	# vowels
train	3696 `sx si`	57110
dev	400 `sx`	6276

Table 2. The data set used in the experiments.

In all experiments mentioned in this paper, three conventional formants are estimated for each vowel by using a formant estimation method proposed by Welling and Ney[11] and followed by a formant tracking algorithm to fix obvious errors. In the classification experiments, the features are the coefficients of 3rd order polynomial approximation of the formant trajectories and the log power. Log duration is also used as one feature[2].

3.1. Classification using perceptual effective formants

To test the possibility of using perceptual effective formants as explicit features in vowel classification, we did a comparison test using both F1' plus F2' and F1 plus F2 as classification features. Results are shown in Table3.

feature	dimension	% correct
E+F1+F2+D	13	66.1%
E+F1+F2'+D	13	62.9%
E+F1'+F2'+D	13	63.5%
E+F1+F2+F3+D	17	70.4%
E+F1+F2'+F3+D	17	68.5%
E+F1'+F2'+F3+D	17	69.2%
E+F1+F2+F3+D+B	20	70.9%
E+F1'+F2'+F3+D+B	20	70.6%
E+F1+F2+F3+D+B+C+P	29	73.2%
E+F1'+F2'+F3+D+B+C+P	29	72.5%

Table 3. Classification comparison using formants and perceptual effective formants. E represent power; D represent log duration; C represent context; B represent average bandwidth; P represent pitch.

The results indicate that there is no advantage to using F1' and F2' as explicit features. There are three possible reasons to explain this result, which seems to contradict the observations in [3] :

1. The estimation of the perceptual effective formants although stable, lost detailed information in low frequency resolution. Therefore the separability between vowels is reduced.
2. The estimation of the perceptual effective formants is not sufficiently accurate to gain the advantages observed using the perceptually measured data.
3. In continuous speech, with contextual effects, the observation of perceptual effective formants and the pure formants might not be as effective as for the isolated vowels like that studied in [3, 6].

To further verify the second possibility and compare it to the assumption made in [3], i.e. F1 and F2' might

separate the vowel space better than F1 and F2, we estimated the first two formants and perceptual effective second formants for the 18 vowels described in Section 2. The results are shown in vowel graphs Figure 1 and Figure 2 as that in [3].

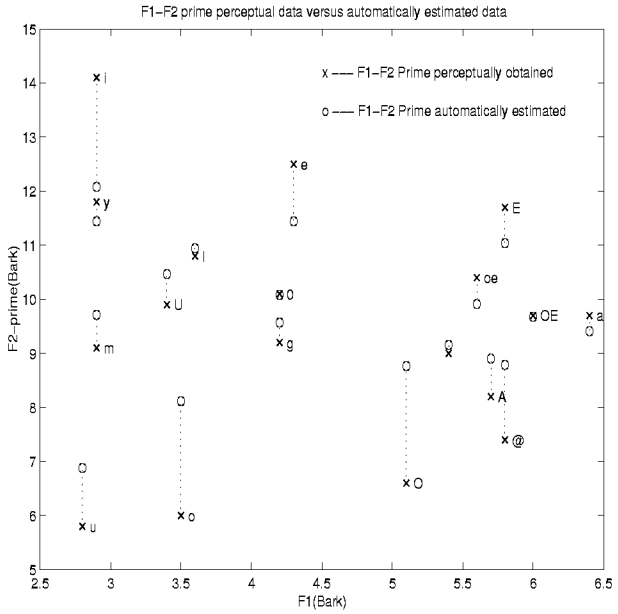

Figure 1. F2' versus F1 of 18 Swedish vowels using perceptual experiment obtained data and automatically estimated data.

Figure 1 shows the vowel diagram (F1-F2') of the perceptually estimated data versus data automatically estimated using our low-order PLP method. It can be seen that on the vowels /o/, /O/ and /i/, the algorithm commits substantial estimation errors which would make the different vowels closer than the perceptual data in the vowel graph. The resulting estimates for these vowels are thus less discriminable than those of the data obtained perceptually.

Figure 2 compares vowel diagrams of F1-F2 versus F1-F2', in which F2' is estimated by the low order PLP method. It shows that becauses of the estimation error of F2', the F1-F2' combinations for these vowels are less separable than the corresponding F1-F2 combinations. This helps to explain the classification results we obtained on TIMIT data.

These plots suggest that, although our estimation of effective formants is quite accurate, the errors committed impact the utility of these features negatively. We next investigate a less direct application of these measurements, which is not sensitive to such fine-scale errors.

3.2. Using perceptual effective formants in formant estimation

As is the case with all formant trackers known to us, our formant tracker sometimes makes gross tracking errors. Several of these errors appear when the energy of F2 is high and the bandwidth is wide. This results in a wide strip of high energy around true F2 position. This might confuse the resonator equations and result in two formant

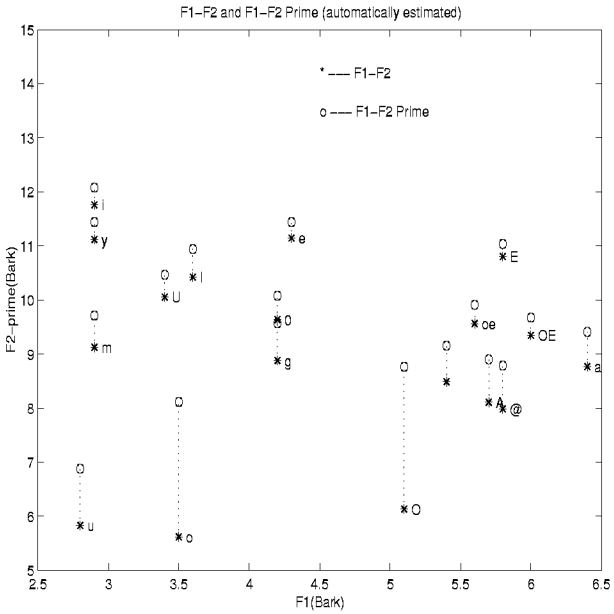

Figure 2. F2 versus F1 and F2' versus F1 of 18 Swedish vowels using automatically estimated data.

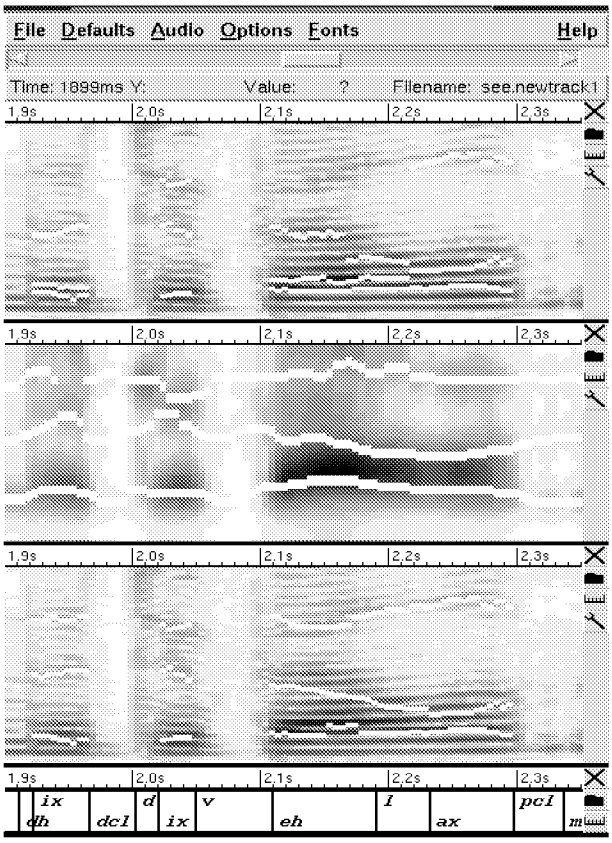

Figure 3. Example of fixing errors using F2' as input.

positions in this broad band while the true F3 is in a much higher frequency band. With the knowledge that F3 should be above F2' in most cases, we implemented an algorithm to check whether a mistake has been made on this we re-estimate the F2 and F3 between $(F1' + F2')/2$ and 4000HZ.[3] This method solved the problem in most of the cases.

One examples is shown in Figure 3. In the figure, the first window shows the original formant estimation result with the formant location (shown in white) on the background of the spectrogram; the second window shows the perceptual effective formant estimations on the background of the PLP spectrogram; the third window shows the improved formant estimation as a result of using F2' as information to correct the errors in the first window where F3 < F2'; the last window shows the TIMIT phoneme labels.

The following tests were performed to evaluate the effect of this post-processing on formant estimation:

1. Manual measurement of errors, in 50 files randomly extracted from the TIMIT training set, which were spoken by 50 different speakers (35 male and 15 female). Formant estimation before and after post-processing were checked by hand. From a total of 772 sonorant segments, the error rate was decreased from 10.2% to 5.7% by post processing (corresponding to a 44.3% reduction in tracking errors) .

2. Vowel classification experiments were performed using formant estimation before this post-processing and after. The results are shown in Table 4:

formant estimation	original	improved
dimension	17	17
% correct	69.8%	70.4%

Table 4. Classification results using different formant estimation methods.

These results show that although the post processing helped reduced some errors in formant estimation, the correction of these errors are not significantly reflected in the classification result.

4. SUMMARY AND FUTURE WORK

We presented an automatic procedure to estimate perceptual second formants. Preliminary experiments show that this estimation method can give relatively good estimates compared to the perceptually measured data. Experiments also show that although perceptual effective formants estimated by this method were not as accurate as conventional formants when used as explicit classification features, they can be used in the formant estimation procedure to help reduce the errors.

We intend to extend this work in a number of directions. First, we would like to understand both why these analysis methods seems to work generally well but produce large error in a few specific cases. We suspect that this can be traced back to the interaction between the sizes and locations of spectral prominences, and will therefore analyze the wide-band spectra in some more detail. Second, we would like to know whether these analysis methods can be adjusted to improve their accuracy in locating F2', therefore improving the separability using F2' directly as a classification feature. In addition, we can also use information of F1' and F2' to correct errors in the estimation of F1 and F2. This would help improve the classification more than correcting F3. In addition, we suspect that a better modeling of the vowels can make better use of perceptual effective formants' information.

Acknowledgement: We sincerely thank Hynek Hermansky for introducing us to this topic, helping with background and many insightful discussions. This work was jointly supported through NSF/DARPA 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.

[3] This method will not hurt the rare cases that F3 <= F2'. In thoses cases, the re-estimated F2 and F3 are the same as the original estimation.

REFERENCES

[1] V. Digilakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 1, pp. 431–442, Oct. 1993.

[2] Z. Hu and E. Barnard, "Smoothness analysis for trajectory features," in *ICASSP-97*, 1997.

[3] G. Fant and A. Risberg, "Auditory matching of vowels with two formant synthetic sounds," *STL-QPRS*, no. 4, pp. 7–11, 1962.

[4] R. Carlson, G. Fant, and B. Granstrom, "Two-formant models, pitch and vowel perception," in *Auditory Analysis amd Perception of Speech* (G. Fant and M. Tatham, eds.), pp. 55–82, Academic, New York, 1975.

[5] R. Carlson, B. Granstrom, and G. Fant, "Some studies concerning perception of isolated vowels," *STL-QPRS*, no. 2-3, pp. 19–35, 1970.

[6] A. Bladon and G. Fant, "A two-formant model and the cardinal vowels," *STL-QPRS*, no. 1, pp. 1–8, 1978.

[7] K. Paliwal, W. Ainsworth, and D. Lindsay, "A study of two-formant models for vowel identification," *Speech Communication*, no. 2, pp. 295–303, 1983.

[8] S. Itahashi and S. Yokoyama, "A formant extraction method utilizing mel scale and equal loudness contour," *STL-QPRS*, no. 4, pp. 17–29, 1978.

[9] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738–1752, 1990.

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

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