Automatic Formant Estimation in a Speech Recognition System

O. Schmidbauer

Siemens AG, ZFE IS KOM31
Otto-Hahn-Ring 6, D-8000 München 83

Abstract

We present an algorithm for formant estimation in continuous speech which is designed to work under "online" conditions in a speech recognition system. The algorithm combines heuristic knowledge about the spectral and temporal behaviour of formants in speech. Preclassification into broad phonetic categories allows to use different algorithms for formant estimation in vowel- and consonant-like regions of speech. Recognition experiments show that formant parameters are a powerful feature set for speech recognition and can compete with other standard feature vectors.

1 Introduction

Formants appear as prominent peaks in the short-time spectra of speech and are defined as the characteristic resonance frequencies of the vocaltract ordered by frequency. Formants carry important information about acoustic-articulatory relations, because they change their frequency and amplitude values according to different vocaltract shapes. They can be viewed as an important source of information in acoustic-phonetic decoding. Thus formants have become a standard in phonetics for describing complex acoustic-phonetic relations.

Formants also seem to be an ideal parameter set for speech recognition, but so far they have not become a standard in this area. The reason is, that automatic formant extraction is not a trivial problem. Already existing algorithms for automatic formant extraction, e.g. [1], [3] show the evidence that formant extraction without any errors is impossible. The significance of information carried by formants is revealed by severe recognition errors in the case of incorrect formant estimation.

The next chapter briefly introduces into the problem of automatic formant extraction. Then the different parts of the algorithm are presented. Finally some speech recognition experiments with formant parameters are described.

2 The Problem

Contrary to commonly used feature sets in speech recognition, formants are *not* defined by a mathematical method, which allows to calculate them directly from the speech wave. They are defined by articulatory phonetics as vocaltract resonances. Formants only can be calculated indirectly via peaks or roots of the power spectrum.

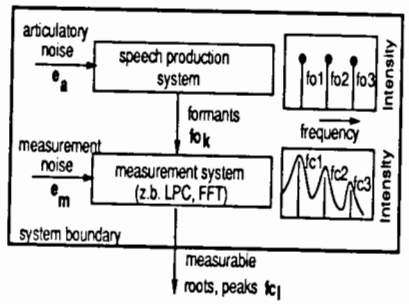

Figure 1: Formant estimation problem.

In terms of estimation theory, we can therefore formulate the following problem (also see figure 1): Suppose the peaks and roots $f\,c_l$ (also called "formant candidates") of the power spectrum are the only data which can be measured and which give us some information about the unknown quantity "formants" $f\,o_k$ inside the system. So, depending on $f\,c_l$ we have to make an estimate for the formants $\hat{f}\,o_k(f\,c_l)$ that the estimation error $E = \hat{f}\,o_k(f\,c_l) - f\,o_k$ is "small".

However, this estimation process is heavily influenced by two different noise sources e_a and e_m: The errors caused by e_a have their origin in the *articulatory* system. The formant order may be confused by zeros in the vocaltract transfer function. Thus some formants are highly damped and are not detectable. Noise source e_m causes *measurement* errors; e.g., as the fundamental frequency is superseeded to the short-time spectra, prominent pitch peaks may be confused with formant candidates.

Existing methods for automatic formant estimation simply try to map these measured and noisy peaks or roots to formants by temporal smoothness criteria, e.g. [1], [3]. The background for these procedures is the assumption that, due to the inertia of the articulators, the temporal behaviour of real vocaltract resonances (=formants) is indicated by continuity.

The algorithm we present in this paper does not exclusively use smoothness criteria, because this is an oversimplification; we will illustrate this point by two examples; firstly, imagine a vowel-segment where a highly damped formant is missing, smoothness criteria do not help at all to classify the measured peaks into formants; secondly, smoothness criteria may lead to crucial errors at places where formants jump significantly in frequency; tracks of different formants may be connected with each other.

3 The Algorithm

Analyzing carefully the temporal and spectral behaviour of formants in speech and also the nature of possible errors we designed an algorithm which can be divided into four steps (see also figure 2): (1) spectral analysis and preclassification into broad categories of manner of articulation, (2) formant identification *(FID)* in vowel-like segments without smoothness criteria, (3) formant tracking *(FTR)* in vowel-consonant (VC) and consonant-vowel (CV) segments with smoothness criteria and (4) preparation and normalization of formant parameters for speech recognition.

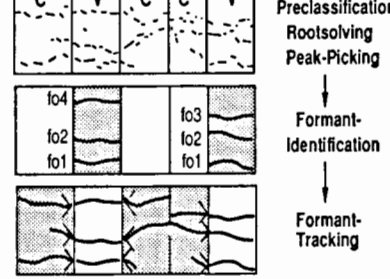

Figure 2: Schematic flow graph for formant-extraction.

The algorithm uses 128-point FFT-spectra with a bandwidth of *8kHz*. The spectra are calculated via a 16-th order *LPC*- analysis with a *20ms* Hamming window, which is shifted in *10m* steps. The formant candidates are determined both by peak-picking and root solving.

3.1 Preclassification

Initially, the speech signal is preclassified into 7 broad phonetic categories *(silence, weak fricative, strong fricative, voiced plosive, nasal, sonorant, vowel)* which correspond to manner of articulation. This is due to the assumption, that there is no overlap of formant frequencies in segments with constant manner of articulation. This makes the following steps of the presented procedure, especially step 2 formant identification, more easily. Classification into categories of manner of articulation is performed by mixture density Hidden Markov Models (CDHMM) similar to [4], using very simple acoustic features like energy contour, zero-crossings rate, low frequency energy (up to 1000 Hz) and the ratio of high to low frequency energy.

3.2 Formant Identification

Formants are extracted in vowel-like (V) segments first, because they usually are more prominent in vowels than in consonants and therefore may be detected more easily. The main task of this step is to allocate formant candidates to formants, taking into account that formants may be missing over the whole duration of a V-segment (see also the example in figure 2). M_{fc} formant candidates are calculated every *10msec*; M_{fc} is set to the number of LPC-roots minus one. Formant identification first tries to find the dominant formant regions within a segment. This is accomplished by approximating the distribution of formant candidates in V-segments by M_{fc} cluster centers with gaussian distributions. The procedure itself consists of three steps: (1) initialization of the cluster procedure, (2) calculation of cluster centers by k-means clustering and (3) classification of the formant candidates into formants by a mean square estimator.

(1) Initialization: To initialize the segment specific formant clusters, we first calculate the mean $m_{f\,c_l}$ and variance $\sigma_{f\,c_l}$ of the formant candidate frequencies $x_{f\,c_l}$ over all N_V frames i of a V-segment:

$$m_{f\,c_l} = \frac{1}{N_V} \sum_{i=1}^{N_V} x_{f\,c_l}(i)$$

$$\sigma_{f\,c_l} = \frac{1}{N_V} \sum_{i=1}^{N} (x_{f\,c_l}(i) - m_{f\,c_l})^2$$

Assuming that we already know $m_{f\,o_k}$, the speaker specific long term formant means $f\,o_k$, the centers of the cluster c_k are initialized to $m_{c_k} = (m_{f\,c_l} + m_{f\,o_k})/2$.

(2) Cluster procedure: The k-means cluster procedure is used to calculate $M_{f\,c}$ cluster centers. The resulting clusters m_{c_k}, ordered by frequency, characterize the segment specific formant frequency regions. They are defined by the mean and variance of M_{c_k} formant candidate frequencies which belong to the $k-th$ cluster:

$$m_{c_k} = \frac{1}{M_{c_k}} \sum_{x_{f\,c_l}(i) \in c_k} x_{f\,c_l}(i)$$

$$\sigma_{c_k} = \frac{1}{M_{c_k}} \sum_{x_{f\,c_l}(i) \in c_k} (x_{f\,c_l}(i) - m_{c_k})^2$$

The speaker-specific formant means $m_{f\,o_k}$ and variances $\sigma_{f\,c_l}$ are calculated by the same cluster procedure, however using a sufficient number of speech frames (about one minute).

(3) Classification: The V-segment specific formant distributions (cluster centers m_{c_k}) are used to classify the formant candidates into formants. The classification procedure maximizes the probability $p(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))$ over all formant candidates $l = 1,..,M_{f\,c}$, i.e. the probability that a measured peak frequency $x_{f\,c_l}(i)$ at time i belongs to formant $f\,o_k$, when it was measured as formant candidate $f\,c_l$. The probability may be written as:

$$p(x_{f\,o_k}(i) \mid x_{f\,c_l}(i)) = \frac{1}{\sqrt{2\pi\sigma_{c_k}}} \exp(-\tfrac{1}{2}(x_{f\,c_l}(i) - m_{c_k})^2 \,/\, \sigma_{c_k}^2)$$

Applying the mean square error criterion [2] to the estimation of formant frequencies leads to the following equation: The estimated frequency $\hat{x}_{f\,o_k}$ of formant $f\,o_k$ is given by the sum of the segment specific mean frequency value m_{c_k} plus the difference of the nearest formant candidate to m_{c_k}, weighted by the maximized probability $p_{max}(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))$:

$$\hat{x}_{f\,o_k}(i) = m_{c_k} + p_{max}(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))\,(x_{f\,c_l}(i) - m_{c_k})$$

3.3 Formant Tracking

This part of the algorithm continues the formants of the vowel-like (V) segments into neighboured consonant-like (C) speech segments, i.e. formant tracking works on CV- and VC-segments. The CV- and VC-segments are well defined by preclassification. As the formants of the V-region are already known, this part of the algorithm has the task to correct and complete corrupted formant tracks by smoothness criteria (see example in figure 2). A nonlinear smoothing algorithm based on dynamic programming was choosen for this task. This method is able to keep frequency jumps in some formants by optimizing the overall smoothness of the formant tracks.

The smoothness of the trajectory of formant f_k is measured by a cost function $c_k(l, i \mid h, i')$. It measures the deviation of formant candidates to the trajectory of formant $f\,o_k$. Assuming that the formant candidates $f\,c_l(i)$ in frame i and $f\,c_h(i')$ in frame i' belong to the trajectory of formant $f\,o_k$ at time i, the costs are given by:

$$c_k(l, i \mid h, i') = \underbrace{[\overbrace{|x_{f\,c_l}(i) - x_{f\,c_h}(i')|}^{1}\, C_1 + \overbrace{(i - i')}^{2} C_2]}_{[p(x_{f\,o_k}(i) \mid x_{f\,c_l}(i))\, p(x_{f\,o_k}(i) \mid x_{f\,c_h}(i')) - 1]}$$

with $i' = i-1, ..i-3$; $\quad l,k = 1,..M_{f\,c}$; C_1 and C_1 being constants.

The cost function consists of three main terms: The first term corresponds to the frequency distance in Hz, the second term measures the temporal distance between the formant candidates and the third one is a weighting term which corresponds to the reverse probability that the formant candidates belong to formant $f\,o_k$. The function accepts small values for smooth and large values for corrupted trajectories.

The optimization criterion for the allocation of formant candidates to formants is given by the next formula. The criterion states that the total error E given by the sum of the costs over all frames N_{VC} for a VC-, N_{CV} for a CV-segment respectively, has to be a minimum:

$$E = min = \sum_{k=1}^{M_{f\,o}} \sum_{i=1}^{N_{VC},N_{CV}} c_k(l, i \mid h, i')$$

over all l, k and i'.

This equation can be elegantly solved by dynamic programming. A solution for this problem is presented in [6].

3.4 Formant Parameters

The formant parameter set which is used for speech recognition consists of 7 formant frequencies and of two energy terms for each formant (a total of 21 parameters). The energy terms correspond to the logarithmic power which is contained in the frequency region extending from a formant center to the left ml or the right minimum mr in the spectrum. With $s(x)$ being the log. power at frequency x, the energy to the left and right and right side $f\,e_{f\,o_k}$ of a formant center is calculated by:

$$f\,e_{f\,o_k} = \int_{x=x_{f\,o_k}}^{x=mr_{f\,o_k}\,(ml_{f\,o_k})} s(x)$$

All formant parameters are finally normalized to the speakers mean values and variances. With $f\,p_k(i)$ now being one the 21 formant parameters at time i and $m_{f\,p_k}$ and $\sigma_{f\,p_k}$ being the speaker specific means and variances of these formant parameters, the normalized formant parameters $f\,n_k(i)$ are calculated by:

$$f\,n_k(i) = \frac{f\,p_k(i) - m_{f\,p_k}}{\sigma_{f\,p_k}}$$

Expressed in filter bank terminology: The resulting parameters which are used for speech recognition are filterbank coefficients, where the filter channels have variable center frequencies and bandwidths.

4 Experimental Results

The presented algorithm for automatic formant extraction was tested with speech material of 3 speakers (each with 2 versions of 100 phonetically balanced sentences, i.e. about 10 minutes of continuously spoken speech per speaker). The extracted formant parameters were used for classifying the speech signal into 14 categories of *place of articulation (silence, glottal, velar, palatal, alveolar, dental-alveolar, labio-dental, bilabial, u-like, o-like, a-like, ö-like, e-like und i-like)*. This task is part of an articulatory based approach for speech recognition [6].

For each articulatory category we built continuous mixture density Hidden Markov models as they are described in [4] and [6]. One version of 100 sentences was used for training, the other version was used for testing. The recognition results on *10ms frame level* are shown in Table 2. The pairs of numbers show the class specific mean recognition rates (left) and the overall frame recognition rates. The formant parameters were compared to a 16-component cepstral vector and to a 64-component feature vector as it is used in [5]. It consists of 32 mel-spectrum coefficients and differential and curvature coefficients, taking into account $\pm 40ms$ of context. The overall mean recognition rate over three speakers (two male, one female) for 21 formant parameters is 74.9 %, for the cepstrum 67.4 % and for the mel-spectrum difference vector 78.5 %. The results show that the formant vector outperforms the cepstral vector (about 7 % better). The recognition performance compared to the the 64-component vector is about 4 percent lower, but it has to be taken into account that the dimensionality of the formant vector is three times lower than for the 64-component vector and that no temporal context was considered for classification.

speaker	21 formant parameters	16 cepstral coefficients	64 mel differential coefficients
male1	74.7 / 84.9	66.8 / 80.3	78.4 / 86.7
male2	74.2 / 84.1	67.3 / 79.5	78.0 / 86.7
female	75.9 / 86.0	68.1 / 81.1	79.2 / 87.9

Table 2: Frame recognition rates [%] for different speakers and different feature sets.

References

[1] P. Laface. A formant tracking system towards automatic recognition of speech. *Signal Processing, North-Holland Publishing Company*, 2:113–129, 1980.

[2] F. Lewis. *Optimal Estimation*. John Wiley and Sons, New York, 1986.

[3] S. McCandless. An algorithm for automatic formant-extraction using linear prediction spectra. *ASSP*, 22:135–141, 1974.

[4] H. Ney and A. Noll. Phoneme modelling using continuous mixture densities. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 437–440, New York, April 1988.

[5] A. Paeseler and H. Ney. Phoneme-based continuous speech recognition results for different language models in the 1000-word spicos system. *Speech Communication, North-Holland Publishing Company*, 7:367–374, 1988.

[6] O. Schmidbauer. *Ein System zur Lauterkennungs auf der Basis Artikulatorischer Merkmale*. Dissertation, Fakultät für Elektro- und Informationstechnik, Technische Universität, München, 1989.