Signal-driven sound processing for uncontrolled environments
Krijnders, Johannes Dirk

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2010

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
8

Automatic extraction of formants in noise


8.1 Introduction

In previous chapters we have shown the application of the recognition techniques developed in chapters 3 and 4 on realistic recordings. While the application to realistic recordings is the main goal of these techniques it is useful to test the techniques on datasets that are carefully recorded and selected because the annotation are less ambiguous. In this chapter the dataset consists of vowels and is annotated both on formants and fundamental frequency. Formants are the resonance frequencies of the vocal tract; they change as the shape of the vocal tract changes. As such, formants are important acoustical cues for the description and identification of phonemes. The task of automatic formant frequency estimation is traditionally investigated by methods based on spectral analysis. Such representations can
be used to accurately estimate the formant positions and formant developments \cite{VargasMcLaughlin2008} in clean speech. However, efforts that focus on formant detection in noise \cite{MustafaBruce2006,HillenbrandHoude2003,deWetHillenbrand2004,YanHoude2006} show much worse performance.

One of the fundamental problems with spectral analysis is that signal and noise are treated alike and spectral shape information is spread over all parameters. As a result, the features are not stable through varying noise conditions. This urges the user to train and test a system in similar conditions. Furthermore the possibility to suppress noise or separate sources after feature estimation is reduced. Improved preprocessing yielded only limited progress towards a solution of this problem. For instance, cepstral mean subtraction can lead to acceptable recognition results \cite{YanHoude2006} but only as long as the acoustic environment complies to highly specific conditions, such as predictable or steady noise. Other methods try to identify unreliable regions before recognition and analyze only the parts that are marked as reliable in order to bias the information towards representing the target speech \cite{CookePhilipp2001}. Such methods work fine as long as enough reliable observations are made which is not the case in SNR's lower than 0dB. Although these and other methods improve the signal descriptions in noisy conditions, the fundamental problem of inclusion of the noise in the spectral features is still unaddressed.

In contrast, human listeners can detect and recognize speech with relatively little hindrance of background noises \cite{Lippmann1997,OShaughnessy2008} which might partly be explained by the characteristics of the extracted features. Human listeners might exploit the fact that some of the informative constituents of the speech sound, namely harmonics near formant positions are relatively robust to noise. Formants or equivalently resonances in the vocal tract stand out energetically and are robust to noise. As a result, the same or similar values could, in principle, be automatically derived from noisy as well as clean speech. In this paper, we test this assumption. We present a newly developed formant-detection algorithm, which can be implemented as a real-time system, that uses features similar to the features hypothesized to be used by humans. We test the robustness of the algorithm to noise and compare with the results from \cite{YanHoude2006}. Next we apply a simple classification method (Best First Search) in order to compare our results with results from \cite{deWetHillenbrand2004} who used the same database to test the robustness of formant-like features.
8.2 Method

8.2.1 Algorithm

To calculate the formant trajectories we calculate a cochleogram of the audio signal with the gamma-tone filterbank as described in section 2.3. For the cochleogram the tone-fit (chapter 3) is calculated and the from the tone-fit matrix signal-components are extracted (chapter 4). An example of signal components extracted from a cochleogram of the utterance “hud” can be seen in figure 8.1(a). These signal components are combined to form harmonic complexes using the methods described in section 4.1.3. The hypothesized harmonic complex with the highest score is used in the next step. The fundamental frequency and selected signal components can be seen in figure 8.1(b). Because not all harmonics are found as signal components the next stages of processing use the energy at the harmonic positions based on the fundamental frequency.

Except for a special case, Lombard speech (section 6.1.1 Junqua (1993)), the formant trajectories do not coincide with the harmonics. Therefore a quadratic interpolation is applied to estimate the real formant location from the harmonics around and including the maximum. This interpolation provides the final estimate of the formant positions as shown in figure 8.1(c). Formant estimates with minimal distance in the frequency plane are connected into formant tracks. Finally we keep formants of sufficient duration (7 frames or more, figure 8.1(d)).

8.2.2 Material

The formant extractor was tested on the American English Vowels dataset (AEV) (Hillenbrand et al., 1994). The dataset consists of 12 vowels pronounced in /h-V-d/ context by 48 female, 45 male and 46 child speakers. All vowels can be correctly classified by American English listeners. The AEV dataset is annotated for the first four formants at 8 points in time for each vowel, which makes it a suitable ground truth. We added pink noise in decreasing signal to noise ratios (SNRs), from 30dB to -14dB SNR. The step size was 10dB at SNR’s above zero and 2dB at lower SNR’s. Pink noise was chosen because it masks speech evenly.

8.2.3 Evaluation

Two performance measures on formant detection for the first three annotated formants are calculated. The first three formants are necessary to classify
Figure 8.1: Cochleogram of a male speaker pronouncing [hud]. (a) Energetic signal components (b) selected HC, the fundamental frequency is given by the striped line (c) formant detections (d) selected formants.

vowels. First, a detection ratio \( r_d \) is calculated, giving the fraction of annotated formants that is consistent with our detections,

\[
r_d = \frac{\#(detected \cap \text{annotated})}{\#(\text{annotated})}
\]  

(8.1)

We consider a detection to be consistent with the annotation if it falls within the range of 15% (1st formant), 12% (2nd formant) and 8% (3rd formant) from the annotated formant frequency. This equals a mean accepted error of respectively 95Hz, 316Hz and 266Hz. The range is chosen such that formants that were considered correct by the authors according to visual inspection were included. Second, a measure is calculated for the detected formants that cannot be related to the annotated formants, the spurious
8.3. Results

Two performance measures on formant detection for the first three annotated formants are calculated. The first three formants are necessary to classify vowels. First, a detection ratio \( r_d \) is calculated, giving the fraction of annotated formants that is consistent with our detections,

\[
r_d = \frac{\#(\text{detected}) - \#(\text{detected} \cap \text{annotated})}{\#(\text{annotated})}
\]

Subsequently, the detected formants that are analogous to the ground truth formants are further investigated in terms of how well they are able to classify the vowels in the test material. To that end, a feature vector is constructed, consisting of the frequency values of only the subset of detected formants that are analogous to the reference formants. Due to missing values, i.e., formants that were not detected, we were limited to a small number of classification algorithms to choose from. The best first tree (BFT) search algorithm from the WEKA toolbox \cite{Witten:2005} allows a weighting of different features. This is a relevant characteristic because different formants represent a different informational value and should be weighted accordingly. We used the BFT search algorithm using a tenfold cross validation method on the detected formants.

8.3.1 Formant extraction

In figure 8.2 (top left and top right, bottom left) the detection rates \( r_d \) and proportion of spurious peaks \( r_{sp} \) are plotted against an increasing SNR. In clean conditions, 90% correct detections are made for all three speaker classes for the first formant, and 75% correct for the second and third formants. For a 0dB SNR, this is reduced for female and child speakers to 70% for the first formant and 53% for the second formant; for male speakers the decline is steeper: for a 0 dB SNR, an \( r_d \) of 50% for the first and 35% for the second formant is found. For all three speaker classes formants consistent with the ground truth can still be extracted at negative SNR values. The proportion of spurious peaks increases gradually to 15% at 0dB for male speakers and -3dB for female and child speakers. To improve the interpretation of the results of the HC extraction stage, table 8.2 shows the occurrences of HC’s that are not detected and the occurrences of HC’s that exhibit an octave error compared to the fundamental frequency annotations in \cite{Hillenbrand:1994}. The percentage of not extracted, as well as wrongly extracted HC’s, is much higher for male speakers than for female and child
3. RESULTS
3.1 Formant extraction

In Figure 2 (top left and top right, bottom left) the detection rates (rd) and proportion of spurious peaks (rp) are shown for a 0 dB SNR. In clean conditions, ~90% of the formants are detected for female (top left), male (top right) and child (bottom left) speakers. For male speakers, the decline is steeper: 3% for the first and 35% for the second formant; for female speakers the decline is less steep: 13% for the first and 41% for the second formant. Detection scores are shown in Table 2. In clean speech, recognition for female speakers is ~80%, for male speakers ~75% and for child speakers ~60%. In 0dB SNR, recognition for female speakers is ~35%, for male speakers ~30% and for child speakers ~25%. For comparison, the BFT search algorithm is shown in the bottom right plane of figure 2. Performance on formant extraction task; percentage of correctly extracted formants according to the annotations for female (top left), male (top right) and child (bottom left) speakers. The bottom right panel gives classification results using a BFT search algorithm.

Figure 8.2: Performance on formant extraction task; percentage of correctly extracted formants according to the annotations for female (top left), male (top right) and child (bottom left) speakers. The bottom right panel gives classification results using a BFT search algorithm.

Speakers which explains the relatively low performance of our method for male speakers. The fact that even in clean conditions some HC's are missed or suffered from octave errors indicates that the criterion on harmonic relations of the tonal components is too strict. Therefore we expect a possible improvement by relaxing this criterion. In the \( r_{sp} \) measure a peak exists at low SNRs for all three speaker classes (figure 8.2 dotted line). This effect is due to an increased number of octave errors in noisy conditions. A fundamental frequency that is too low results in more harmonics between two formant positions which explains the relatively high amount of spurious peaks. If the SNR decreases, the number of incorrectly extracted harmonic complexes increases, which results in an increased amount of incorrect formant detections. If the SNR decreases further, the overall number of extracted harmonic complexes decreases due to missed harmonic complexes.
8.3.2 Vowel classification

The bottom right panel of figure 8.2 shows the classification scores obtained with the BFT search algorithm. Recognition in clean speech is 80% for all three speaker classes. In 0dB SNR, recognition for female speakers is 60% and recognition for male speakers 35%. Table 8.1 shows the confusion matrix of the classifications of the vowels from all speakers pooled together. Relatively many confusions occur between the vowel sounds ‘ae’, ‘eh’ and ‘ah’, ‘aw’. Those four vowels are confused with one of the other sounds for 25% percent of the vowels. It is noteworthy that the same vowels are reported to be confused most often by human listeners [Hillenbrand et al., 1994], see table 8.1.

8.4 Discussion

We described and tested a method to automatically extract formants based on robust parts of the acoustical signal, namely the harmonics at formant positions. In contrast to commonly used ASR features that degrade slowly as a function of decreasing SNR, formant positions remain constant under different noise conditions. The robustness of harmonics at formant positions allows us to develop a method to extract similar feature values over a range of acoustical conditions.

8.4.1 Harmonic complex extraction

The results in table 8.2 show that in clean situation the harmonic complex extraction works near perfect for female and child voices. This performance only starts to drop around 0 dB. The scoring function (equation 4.1) was not
Table 8.2: Type of mismatch for detection of the harmonic complex for male, female and child speakers. For male speakers more harmonic complexes are missed and more octave errors are made.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>30</th>
<th>20</th>
<th>10</th>
<th>0</th>
<th>-2</th>
<th>-4</th>
<th>-6</th>
<th>-14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female Not extracted(%)</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>18</td>
<td>23</td>
<td>35</td>
<td>51</td>
<td>98</td>
</tr>
<tr>
<td>Octave error(%)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>10</td>
<td>11</td>
<td>13</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Male Not extracted(%)</td>
<td>2</td>
<td>1</td>
<td>8</td>
<td>41</td>
<td>74</td>
<td>81</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Octave error(%)</td>
<td>8</td>
<td>6</td>
<td>10</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Child Not extracted(%)</td>
<td>0</td>
<td>2</td>
<td>17</td>
<td>28</td>
<td>39</td>
<td>51</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Octave error(%)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

optimized for this dataset, but rather for the dataset used in chapter 4. This leaves room for improvement. The performance for male speakers, with a lower fundamental frequency, leaves room for future work as well.

8.4.2 Formant extraction

In noisy conditions our method compares well with existing methods proposed in the literature. For instance the method proposed by (Yan et al., 2006) is a linear prediction model consisting of a noise reduction stage, a secondary hidden Markov model (HMM2) and a Kalman filter. This method results in average estimation errors of respectively 17%, 12% and 8% for the first, second and third formants in a SNR of 0dB train noise. Although the method in (Yan et al., 2006) outperforms our method, a problem of the method proposed by (Yan et al., 2006) is that it cannot be easily generalized to unseen types of noise, as the noise reduction methods are specifically suitable for relatively stable types of noise and the method relies predominantly on de-noising of the input signal. In contrast to this our method can be generalized to all types of noise.

8.4.3 Vowel classification

In noisy conditions the results for female and child speakers compare favorably with results found in the literature. (de Wet et al., 2004) report on a vowel classification task on the same AEV database, in which they used HMM2 to evaluate probabilities of both frequency and time. Using this method 55% correct classifications in 0dB babble noise are found for female and male speakers. For female speakers our method results in higher scores (60%) although we used a simple classification mechanism (BFT search). We already mentioned the possibility of improving the part of the algorithm that extracts the harmonic complex and it is possible that identification can be improved with a more advanced learning algorithm. This yields the possibility to obtain better results in noisy conditions for all speaker classes compared to those reported by (de Wet et al., 2004). Apart from this positive
comparison, it is noteworthy that by using features similar to the features hypothesized to be used by humans, we find confusions similar to those of human listeners.

8.5 Conclusion

We showed that it is possible to develop an automatic method to extract formant feature values over a range of acoustical conditions that uses the robustness of harmonics at formant positions. For pink noise we showed that formants consistent with the ground truth can be extracted at low and even negative SNR-values. We expect performance enhancement by further optimizing our harmonic complex detection algorithm. These initial results seem to suggest that formants, thought to be important for humans in speech processing, can also constitute robust features for automatic vowel detection, and possible automatic speech recognition, systems.