Psycho-acoustically motivated formant feature extraction

Bea Valkenier
University of Groningen
Groningen, the Netherlands
b.valkenier@ai.rug.nl

Dirkjan Krijnders
University of Groningen
Groningen, the Netherlands
j.d.krijnders@ai.rug.nl

Ronald A.J. van Elburg[1]
University of Groningen
Groningen, the Netherlands
RonaldAJ@vanElburg.eu

Tjeerd C. Andringa[1]
University of Groningen
Groningen, the Netherlands
t.c.andringa@ai.rug.nl

[1]These authors contributed equally.

Abstract

Psycho-acoustical research investigates how human listeners are able to separate sounds that stem from different sources. This ability might be one of the reasons that human speech processing is robust to noise but methods that exploit this are, to our knowledge, not used in systems for automatic formant extraction or in modern speech recognition systems. Therefore we investigate the possibility to use harmonics that are consistent with a harmonic complex as the basis for a robust formant extraction algorithm. With this new method we aim to overcome limitations of most modern automatic speech recognition systems by taking advantage of the robustness of harmonics at formant positions. We tested the effectiveness of our formant detection algorithm on Hillenbrand's annotated American English Vowels dataset and found that in pink noise the results are competitive with existing systems. Furthermore, our method needs no training and is implementable as a realtime system which contrasts many of the existing systems.

1 Introduction

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 LPC. Such representations accurately estimate formant positions and formant developments (Vargas and McLaughlin, 2008) in clean speech. However, efforts that focus on formant detection in noise (de Wet et al. 2004; Mustafa and Bruce, 2006; Yan et al. 2007) show results that deteriorate quickly in noise. One exception to this can be found by the system that was recently developed by Glaeser et al. (2010); their method shows a major improvement with regard to other methods.

Human listeners can detect and recognize speech in uncontrolled environments with relatively little hindrance of background noises (O'Shaughnessy, 2008). Psycho-acoustical research suggests that human listeners use Bregmans grouping cues (Bregman, 1990) to recombine components of sounds into a single percept. Provided the individual components are separable from background noise these grouping principles can be applied in automatic methods. Those methods were first investigated by Duifhuis et al. (1982). In general, systems based on grouping of harmonics are applicable in uncontrolled environments and do not rely on training. However, harmonic mismatches or missed detections sometimes occur.

Here, we investigate whether we can use the extractions of a harmonic grouping algorithm to extract robust formants without the need of training. Our results show that formant position estimates are stable over different noise conditions for a simple database. The results indicate that a renewed investigation of the problem of harmonic complex extraction can be a key to solving the lack of robustness in features for applications such as automatic speech recognition.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 218–223

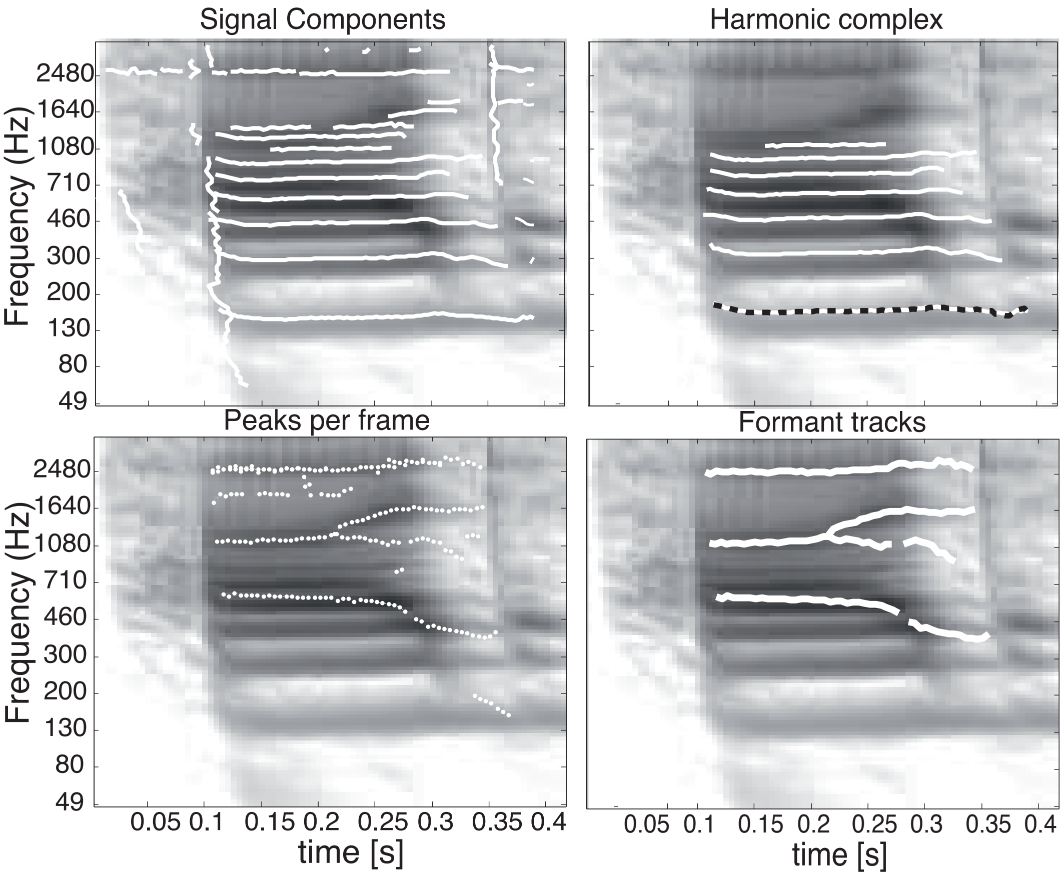

Figure 1: *Results of the different steps in the algorithm represented on a cochleogram of a male speaker pronouncing [hud]. (top left) Energetic signal components; (top right) selected HC, the fundamental frequency is given by the dashed line; (bottom left) formant detections based on this fundamental frequency and its overtones that fall below 4000Hz; (bottom right) selected formants*

2 Methods

2.1 Algorithm

In order to reach a close estimation of the resonance frequencies of the vocal tract we perform peak interpolation over harmonics in a harmonic complex (HC). First, the time signal is converted to the time-frequency domain by a gamma-chirp filterbank (Irino and Patterson, 1997). Its filter coefficients (h_{gc}) are defined by,

$$h_{gc} = at^{N-1}e^{-2\pi bB(f_c)t}e^{j(2\pi f_c t + c\log(t))} \qquad (1)$$

where N = 4 is the order of the gammachirp. The coefficients (a = 1, b = 0.71, c = -3.7) are based on Irino and Patterson (1997) but were adjusted such that the response is narrower in frequency such that the tonal components become emphasized. The frequency range f_c is fully logarithmic from 67 to 4000 Hz over 100 channels. The bandwidth (B) of the filters is given by (9),

$$B(f_c) = 24.7 + 0.108 f_c \qquad (2)$$

We call the averaged and logarithmically compressed result a cochleogram.

Second, harmonics are extracted from the cochleogram using tone fit. Here we only give a global description of tone fit (see Krijnders and Andringa (submitted) and Krijnders et al. (2009) for details). The tone fit is a measure how well the cochleogram matches a tone at that time-frequency location. This measure is calculated with a filter derived from the response of the cochleogram to a perfect tone. Connected locations that match the filter well (with high tone-fit values) are extracted and are described as a line through the best matching location. We call such a description a signal component (Figure 1, top left).

The final step before the formant extraction combines signal components into HCs (Figure 1, top right). To that end, HC hypotheses are generated from energetic signal components (Figure 1, top right) that partly overlap in time and have an approximately harmonic frequency relation to each other. Initially a hypothesis consists of a fundamental frequency (f_0) estimate and energetic signal components. Additional signal components are added later to each hypothesis if they increase the score of that hypothesis. This score is defined as (Krijnders et al., 2009; Niessen et al., 2009):

$$S = n_{sc} + b_{f0} + n_h - \sum_{sc} rms_{sc} - \sum_{sc} \Delta f_{sc} \quad (3)$$

where n_{sc} is the number of signal components in the group, b_{f0} is one or zero depending on the existence of a signal component at the f_0, n_h is the number of sequential harmonics in the group, rms_{sc} are the root mean square values of the differences of the signal component f_0 after the mean frequency difference is removed, and Δf_{sc} is the mean frequency difference divided by harmonic number. To reduce octave errors additional hypotheses at octaves above and below each hypothesis are added and scored. In the formant extraction phase only the hypothesis with the highest score is used.

The resonance frequencies of the vocal tract might be located between two harmonics. Therefore, a three point quadratic interpolation over the harmonics around the harmonic with (local) maximum energy is used to estimate the formant location (Figure 1, bottom left). Subsequently, formant estimates with minimal distance in the adjacent frames in the time-frequency plane are connected into formant tracks. Only tracks of sufficient duration (7 frames or more, Figure 1, bottom right) are kept. These long formant tracks constitute our final formant estimate.

2.2 Material

The formant extractor was tested on the American English Vowels dataset (AEV) HillenBrand (1995). The dataset consists of 12 vowels pronounced in /h-V-d/ context by 48 female, 45 male and 46 child speakers. The AEV dataset is automatically annotated and subsequently hand-corrected 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 -6dB SNR. Pink noise was chosen because it masks speech evenly.

2.3 Evaluation

As we do not extract exactly three formants we cannot calculate error scores that represent the distance of the extracted formant to the annotated formant. The annotations are determined in clean speech and therefore we compare our results to the annotations for the clean speech condition. In order to evaluate the robustness of the system, we compare our results in noise to our results in clean speech as this gives the best estimate of noise robustness of the features.

2.3.1 Detections in clean speech

We specify two performance measures that together indicate how useful the features are for classification and calculate those for the features extracted in clean speech. The usefulness for classification is based on extraction of informative features on the one hand and neglecting non-informative features on the other hand.

The r_d gives the fraction of annotated formants that is consistent with our detections,

$$r_d = \frac{\#detected \cap \#annotated}{\#annotated} \quad (4)$$

We consider a detection to be consistent with the annotation if the relative error falls within the range of 15% (1st formant), 12% (2nd formant) and 8% (3rd formant). 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.

The ratio spurious peaks (r_{sp}) is a measure for the detected formants that cannot be related to the annotated formants. It gives the ratio between the number of extra detected formants at the annotated positions, and the number of annotated points,

$$r_{sp} = \frac{\#detected - (\#detected \cap \#annotated)}{\#annotated} \quad (5)$$

The r_d and the r_{sp} are used to compare our results in clean speech to the annotations of the database. The robustness of the features is not determined with regard to the annotations but by comparing the results to itself.

2.3.2 Precision and recall in noise

The robustness is calculated by the precision and recall of the findings in clean speech. The precision reflects whether the extracted formants are

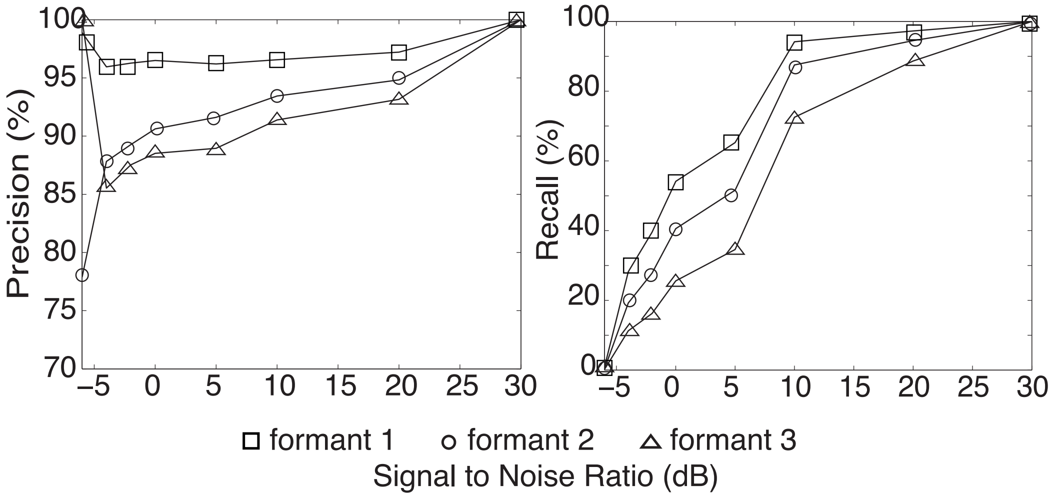

Figure 2: *Left panel: Percentage of correctly extracted formants (i.e. relative error falls within the range of 15% (1st formant) 12% (2nd formant) and 8% (3rd formant) in increasing SNR levels in pink noise. Right panel: Percentage not detected formants in increasing SNR levels in pink noise.*

relevant (with regard to the ground truth), it gives the amount of correct detections relative to the total amount of detections for a noise condition.

$$precision = \frac{\#truepositive}{\#truepositive + \#falsepositive} \quad (6)$$

The recall gives the amount of not detected formants relative to the total amount of detections for a noise condition reflecting whether the formants in the ground truth are extracted in noise as well.

$$recall = \frac{\#truepositive}{\#truepositive + \#falsenegative} \quad (7)$$

3 Results

3.1 Detections in clean speech

The detection rates (r_d) and proportion of spurious peaks (r_{sp}) are calculated for clean speech with regard to the annotated formants. 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. The level of spurious peaks is found at 10%.

Table 1: *Type of mismatch for detection of the harmonic complex for male, female and child speakers in pink noise. For male speakers more harmonic complexes are missed and more octave errors are made.*

	SNR(EdB)	30	10	0	-4	-6
female	not extracted	0	1	18	35	51
	octave error	1	3	10	13	11
male	not extracted	2	8	41	74	81
	octave error	8	10	7	3	3
child	not extracted	0	1	17	39	51
	octave error		2	4	9	8

3.2 Precision in noise

In Figure 2 the precision of the findings is plotted against an increasing SNR in pink noise (left panel). Formants consistent with the ground truth can still be extracted at negative SNR values. Performance stays very high for the first formant and remains above 75% for both the second and the third formant.

3.3 Recall in noise

The right panel in Figure 2 shows that the recall is high above 10dB SNR and decreases rapidly in higher noise levels. The main reason for this is that harmonic complexes are not, or not correctly extracted. To provide a better insight in the results of the HC extraction stage, table 1 shows the occurrences of HCs that are not detected and the occurrences of HCs that exhibit an octave error in pink noise, calculated on the f0 annotations in Hillenbrand (1995).

4 Discussion

We described and tested a method to automatically extract formants based on robust parts of the acoustic signal, namely the harmonics at formant positions. The robustness of harmonics at formant positions allows us to develop a method to extract similar feature values in varying SNR. Because the harmonics have high energy levels the influence of noise is relatively small. The energetic harmonics provide a solid bases for the extraction of formants that are important acoustical cues for the identification of phonemes. With the aim of developing a system for robust phoneme identification speech features derived from harmonics are a good starting position. We showed that it is possible to extract formant feature values over SNRs from 30dB to -6dB in pink noise, that uses the robustness of harmonics at formant positions in human hearing. These initial results support the believe that harmonic grouping can be used as a basis for speech processing.

Recently Glaeser et al. (2010) presented a method that robustly estimates formant positions. In 0dB SNR they find mean relative error scores of approximately 24%, 17% and 10%, which is slightly worse than our results probably because it was tested on a more challenging database. One important difference of their method is that it is based on the enhancement of harmonics instead of grouping. We expect therefore that our method is better suited for data with mixed sources such as competing speakers.

Because the extraction of harmonic complexes poses some unsolved problems such as misses and octave errors we argue that the problem of the extraction of harmonic complexes should be systematically investigated. If this problem can be solved we have access to extremely robust features for speech coding with the advantage that training on a specific noise condition is not needed.

4.1 Conclusion

We showed that it is possible to develop an automatic method to extract formant feature values over SNRs from 30dB to -6dB in pink noise, that uses the robustness of harmonics at formant positions in human hearing. These initial results support the believe that harmonic grouping can be used as a basis for speech processing.

5 Acknowledgements

BV was supported by STW grant DTF 7459, JDK was supported by NWO grant 634.000.432. The authors would like to thank Odette Scharenborg, Jennifer Spenader, Maria Niessen, Hedde van de Vooren and three anonymous reviewers for their useful comments on earlier versions of this manuscript.

References

Bregman, A.S., “Auditory scene analysis: The perceptual organization of sound”, Cambridge, Massachusetts: The MIT Press 1990.

Glaeser,C., Heckmann, M., Joublin, F. and Goerick,C. “Combining auditory preprocessing and bayesian estimation for robust formant tracking,” IEEE trans. on audio, speech and language processing, 18(2), pp. 224 - 236, 2010

Hillenbrand, J.M., Getty, L.A., Clark, M.J. and Wheeler, K. “Acoustic characteristics of American English vowels,” J Acoust Soc Am, 97, pp. 3099 - 3111, 1995

Irino, T. and Patterson, R.D. “A time-domain, level-dependent auditory filter: The gammachirp,” J Acoust Soc Am, 101(1), pp. 412 - 419, Jan 1997.

Krijnders, J.D., Niessen, M.E. and Andringa,T.C. “Sound event recognition through expectancy-based evaluation of signal- driven hypotheses,” Pattern Recognition Letters, accepted 2009.

Krijnders, J.D. and Andringa, T.C. “Tone, pulse, and chirp decomposition for environmental sound analysis,” Submitted

Moore, B.C.J. “A revision of Zwicker’s loudness modell.” Acustica, 82(2), pp. 335 - 345(11), 1996.

Mustafa, K. and Bruce, I.C. “Robust Formant Tracking for Continuous Speech With Speaker Variability,” IEEE trans. on audio, speech and language processing, 14(2), pp. 435 - 444, 2006

Niessen, M., Krijnders, J.D., and Andringa, T.C. “Understanding a soundscape through its components”. Proceedings of Euronoise 2009

O’Shaughnessy, D. “Invited paper: Automatic speech recognition: History, methods and challenges,” Pattern Recognition 41(10), 2965 - 2979, 2008

Vargas, J. and McLaughlin, S.“Cascade Prediction Filters With Adaptive Zeros to Track the Time-Varying Resonances of the Vocal Tract,” IEEE trans. on audio, speech and language processing, 16(1), pp. 1 - 7, 2008

Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S. and Bourlard, H. “Evaluation of formant-like features on an automatic vowel classification task,” J Acoust Soc Am 116, pp. 1781 - 1791, 2004.

Yan, Q., Vesghi, S., Zavarehei, E., Milner,B., Darch,J., White, P. and Andrianakis, I. “Formant tracking linear prediction model using HMMs and Kalman filters for noisy speech processing,” Computer speech and language 21, pp. 543 - 561, 2007