Robust Vowel Detection

B.Valkenier[1], J.D.Krijnders[1], R.A.J. van Elburg[1,*] & T.C. Andringa[1,*]

[1] *University of Groningen, The Netherlands,*

** These authors contributed equally to this work*

Email: <B.Valkenier; J.D.Krijnders; R.A.J.van.Elburg; T.C.Andringa>@ai.rug.nl

Introduction

Human listeners can detect and recognize speech with relatively little hindrance of background noises [1]. It is well known that human listeners apply knowledge to derive a coherent interpretation of ambiguous input. But human listeners might also benefit from the robustness of formant patterns. Formants are part of the perceptual features that are hypothesized to be used by humans in speech processing [2, 3, 4] and correspond to the resonance frequencies of the vocal tract. Shape changes of the vocal tract influence the formants and lead to their development over time. The temporal development of formants is generally thought to characterize vowels. These developing formant patterns, are noticeable in the acoustic signal as amplified energy of the harmonics of the speech sound and are as such robust to noise.

While perceptual features focus on spectral detail, the representations of speech used in Automatic Speech Recognition (ASR) focus on describing the spectral envelope. For the special case of clean speech such ASR-features can be used to accurately determine the formant positions and formant developments [5]. But efforts that focus at formant detection in noise [6, 7, 8] still perform well below human performance. One of the reasons for the relatively low performances in noise might be that ASR-features treat signal and noise alike and spread spectral shape information over multiple parameters. As a result of this, the possibility to suppress noise or separate sources after feature estimation is reduced. Advances have been made to deal with this problem. First, the effects of noise can be separated from the target speech in some special cases. For example, in the case of stationary noise it is possible to remove a constant noise component in the features with cepstral mean substraction. Such methods can lead to acceptable recognition results in highly specific conditions [9]. Second, methods exist that ignore unreliable features and as such bias the information towards representing the target speech [10]. Although those, and other, advances have been made that improve the signal descriptions in noisy conditions, the fundamental problem has not been solved. This implies that the remaining noise must be dealt with during the pattern recognition phase. Currently this is only possible by tuning the ASR-system for specialized applications which allow the input to be from rather narrow and preferably constant domains.

The fact that human listeners still outperform ASR systems in noise [1, 11] might partly be explained by the different characteristics of the extracted features. In contrast to whole spectral shape features, perceptual features are robust to noise. As a result of this robustness, the same or similar feature values will be derived from noisy as well as clean conditions. To pursue this idea further we developed a formant-detection algorithm using features similar to the features hypothesized to be used by humans.

The aim of this study is twofold. First we determine whether the subset of extracted formants includes the reference formants, and second we test whether those can be used to classify vowels correctly. We test the performance of the method neither optimizing the detection nor the classification method and using a preliminary experimental test. In the method section we explain the algorithm and experiment to test our method. We report on formant detection and vowel classification performances in the result section. We focus our discussion on the performance in noise because this is where, even without optimizing the algorithm we expect comparative advantages of our method.

Method

The features used in our formant-detection algorithm are derived from the speech signal in several steps whose results are illustrated in Figure 1. First, we compute a cochleogram and identify in it high energy regions of suitable shape and sufficient size (a). Then we extract possible harmonic complexes (HCs), and complement them with less reliable signal components (b). In the next step we determine local formant positions by establishing the maxima in an polynomial interpolation between peaks in the harmonic complex(c). Finally we select formants of sufficient duration (d). The algorithm is not optimized for usage on specific speech sounds or this specific database.

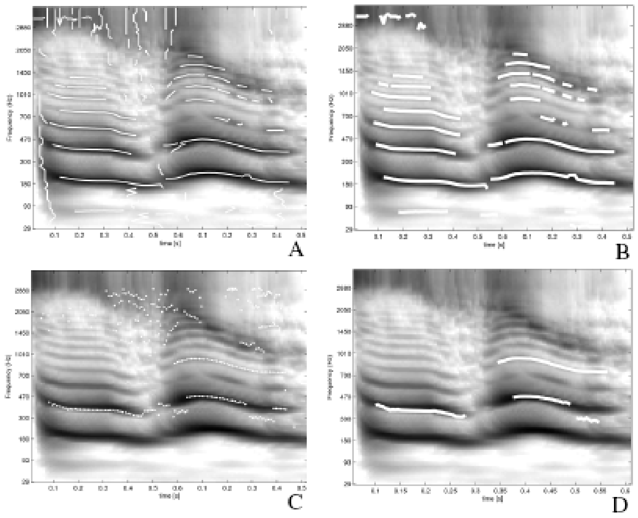

Figure 1: Cochleogram of the utterance [ubi] in clean speech. Dark grey regions depict high energy levels, light grey regions depict low energy levels. (a) energetic signal components (b) selected harmonic complexes (c) formant detections (d) selected formants

We applied our method to the American English Vowels dataset (AEV) [12] with added pink noise in decreasing signal to noise ratios (SNRs). Pink noise was chosen because it masks speech evenly. 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 at the level of formants at 8 points in time, which makes it a suitable ground truth.

From the detections two performance measures on formant detection for the first three annotated formants are calculated. First, a detection ratio (r_d) is calculated, giving the fraction of annotated formants that is consistent with our detections,

$$r_d = \frac{\#\{\text{detected} \cap \text{annotated}\}}{\#\{\text{annotated}\}}.$$

We consider a detection to be consistent with the annotation if it falls within two standard deviations from the mean of the reference formant of a class as obtained from the annotations. Second, a measure is calculated for the detected formants that cannot be related to the annotated formants, the spurious peaks (r_{sp}). This measure is the ratio between the number of extra detected formants at the annotated positions, and the number of annotated points,

$$r_{sp} = \frac{\#\{\text{detected}\} - \#\{\text{detected} \cap \text{annotated}\}}{\#\{\text{annotated}\}}.$$

In addition, we investigated how well the detected formants that are analogous to the ground truth formants can be used to classify the sounds according to vowel quality (class) without making any claims regarding the machine learning algorithm we used. Hereto the best first tree search algorithm from the WEKA toolbox is used [13]. 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 tree search algorithm allows a weighting of different features. This is a relevant characteristic because different formants represent a different informational value and should be weighted accordingly. The best first tree search algorithm used a ten-fold cross validation method on both the detected formants and on the ground truth formants.

Results

In Figure 2 the detection rate (r_d, solid line) and the proportion of spurious peaks (r_{sp}, dotted line) are plotted against an increasing SNR. Separate results are plotted for female (a), male (b) and child speakers (c). Mean overall results are plotted in figure 2 (d). The overall detection performance is around 65% for clean conditions and decreases to 35% in a SNR of 0dB down to 0% in a SNR of -15dB. The proportion correct detections are higher (~75%) for female (a) and child (c) speakers than they are for male speakers (~50%). Also the proportion of spurious peaks is lower (~20%) for both female and child speakers than for male speakers (~30%).

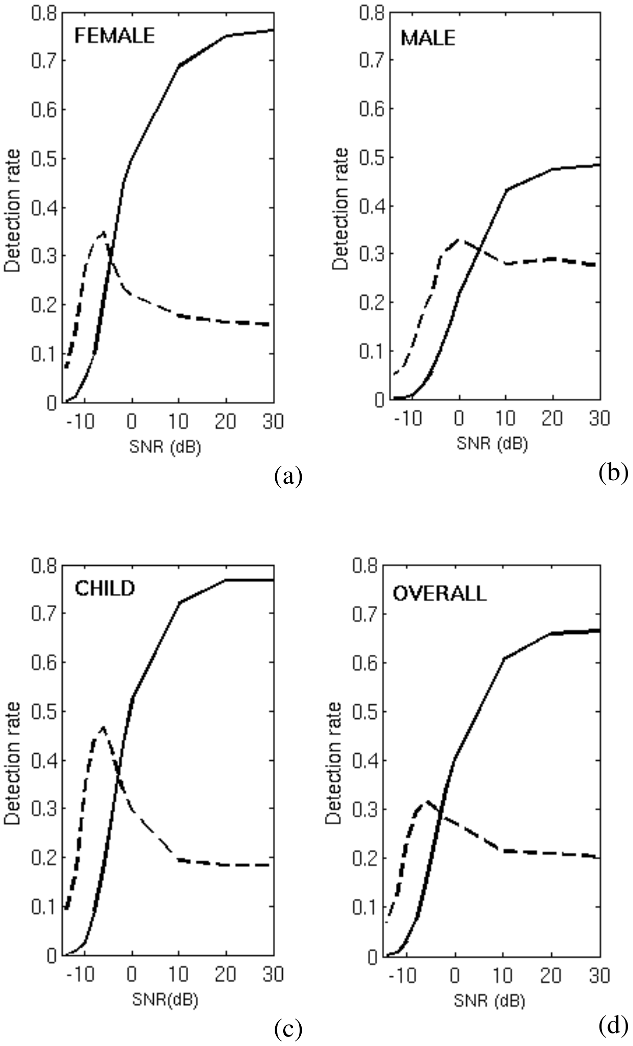

Figure 2: (a,b,c,d) Proportion of the annotated formants [12] that is covered by our subset of detected formants (r_d, solid). Proportion extra detected formants, spurious peaks, in our subset (r_{sp}, striped). (a) women (b) men (c) children (d) pooled results over a,b and c

In Figure 3 the classification results are plotted against an increasing signal to noise ratio. The classification performances are slightly higher than the detection performances (Figure 2). Based on the formant detections (between 50% and 75%), a classification score between 65% and 75% is reached in clean conditions. For a SNR of 0dB the female as well as child speaker performance is improved to 55% correct classification. The classification result based on the ground truth formants shows 88% correct classifications.

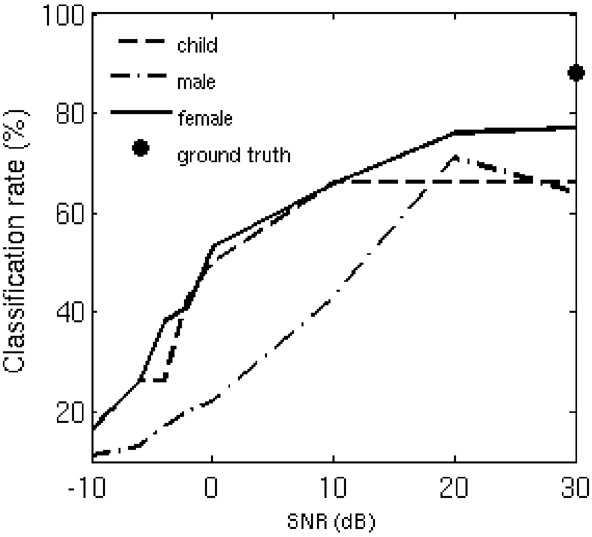

Figure 3: Percentage correctly classified vowels by means of a best first search for the formants as annotated by [12] (dot) and for the subset of detected formants that covers the annotations for SNR -10dB up to SNR 30dB for female (solid) child (striped) and male speakers (dotted and striped).

Discussion

In this article we described and tested a method to automatically extract formants based on the notion of perceptual features. In contrast to commonly used ASR-features perceptual features remain similar throughout different noise conditions. This characteristic can be exploited by our method as long as the features are not masked by noise. The robustness of perceptual features allows us to develop a single method to extract similar feature values through varying acoustical conditions. In pink noise we showed that formants consistent with the ground truth are extracted into the low SNR range, and can be used to classify vowels.

Three things attract our attention if we have a closer look at figure 2. First, in an SNR of 0dB pink noise our method still detects 50% of the correct formants for female speaker and 55% for child speakers but only 25% for male speakers. We cannot yet explain this big difference between the results for male speakers and child and female speakers, but it is at least in part due to imperfections in our algorithm for finding the correct harmonic complex rather than a failure of our features. Second, a peak exists at low SNRs for all three speaker classes in the r_{sp} measure, and is less pronounced for the case of male speakers than it is for the child and female speakers. A similar explanation can be given for this observation. If the SNR decreases, the number of incorrectly extracted harmonic complexes will first increase, resulting in an increased amount of incorrect formant detections. Then, if SNR decreases further, the number of incorrectly extracted HCs will decrease again because no HC is extracted anymore. In the case of male speakers there is an increased amount of incorrectly extracted HCs from the beginning which explains the relatively low peak for this speaker class. Third, the performance in clean conditions are relatively low compared to other methods [8]. An explanation for this might be that the detected harmonic complex does not always perfectly coincide with the points at which the utterance is annotated. Therefore, the first annotations for all three formants are often missed. The r_d measure gives credit for all annotated points where we find a formant that is annotated. If the detected harmonic complex is time shifted and is therefore not properly aligned with the annotations a relatively low score for those sound-files is obtained.

Two possible effects of choices we made do not directly follow from the results. First, the extracted harmonic complexes are often much longer than the annotations and this is not credited although visual inspection indicates that the extracted formants are still correct. This implies that the measure r_d might not be optimal. However, because data that is annotated on the level of formants is scarce, we cannot easily switch to another measure or database. Second, we used a tolerance of two standard deviations of the mean reference formants, averaged over all vowel qualities. This choice is based on the idea that two standard deviations is a naturally occurring variability of formants, in which humans are still able to classify the vowel input. One of the adverse effects of this choice might be that not all detected formants are reported. Expectedly, this effect is more prominent in noise than in clean conditions because additive noise might result in increasing standard deviations whereas the tolerance is calculated on the ground truth formants in clean speech. If a useful selection or classification algorithm is implemented this problem will be solved because a tolerance value is not needed anymore.

We like to compare our results to two recent studies reporting formant detection scores in noise. At present a direct quantitative comparison with those methods cannot be made due to different performance measures and noise conditions.. Therefore we describe in qualitative terms both the method and the obtained results in noise. The method proposed by [6] uses a prefiltering technique where the speech is filtered by a time varying adaptive bandpass filter before formant frequency estimation. Effectively the application of this method results in the dynamic amplification of interesting frequency regions. In 0dB white noise the method leads to performance errors close to the standard deviation of the formant frequencies over the whole utterance. This result can likely be explained by a regression towards the filter respons by decreasing SNR. This means that by increasing noise levels, the derived formant frequency gradually changes towards the filter respons and therefore the extracted formant features are not reliable in noise. The method proposed by [9] is a three stage formant tracking linear prediction model. In the first stage a noise reduction is performed. In the second stage a secondary hidden markov model (HMM2) is used to track formant variation in both time and frequency, and in the final stage a Kalman filter is used to give a smoothed trajectory. 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. One of the problems of this method is that it is not clear how it can be generalized to other, less predictable types of noise. The noise reduction methods used are specifically suitable for relatively stable types of noise and the method relies for a big part on denoising of the input signal.

Besides applying a machine learning algorithm to our results we also used the reference formants from the AEV dataset to test the classification method. Using the best first tree algorithm from the Weka toolbox [13] we find a classification performance of 88%. Using the same dataset and therefore the same features [7] classified around 96% correct using a weighted narrow-band pattern matching method. From the clear difference between those classification results (96%) and our result (88%) we conclude that the machine learning algorithm we used is not optimal. Another classification mechanism might therefore give better classification results, not only in clean, but probably also in noisy conditions. However, we applied a classification method in order to determine whether the extracted formants can be used to classify vowels in changing noise conditions.

In an SNR of 0dB we find a vowel classification performance of 55% (figure 3) for female and child speakers based on an overall formant detection rate of 45%. A similar classification result in 0dB babble noise is found by [8] on the same AEV database. They use 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. An important characteristic of this method is that it uses statistical, database specific knowledge (in the form of frequency probabilities) to select the correct formants from a set of detected formants. Therefore [8] might profit from the relatively stable conditions of this database.

We expect a possible improvement in detection scores by incorporating knowledge in a decision mechanism. One of the possible improvements is that in the selection stage incorporating knowledge might lead to a reduction of the amount of spurious peaks by selecting those formants that lead to a consistent vowel hypothesis. Another possible improvement might be found in the classification stage. Provided still some information on vowel quality can be estimated, a knowledge guided search can converge to a single percept, even if the signal is heavily degraded. Encouraging examples for such approaches can be found in technical as well as psycholinguistic literature. First, the missing features method [10] is a method where unreliable features are ignored and subsequently a coherent classification is attempted with the remaining features. Second, from a psycholinguistic point of view convergence onto a single percept can result from a competition between different possible percepts. In [2, 3, 4] the effect of such a competition is that the input that matches knowledge best receives the highest activation and as such results in a single percept.

In this article we showed that our method allows us to extract robust features in varying acoustical conditions. Until fairly low SNRs the extracted features can still be used to correctly classify vowels. We still expect strong performance improvement with further improvements on our method. Especially through improvements in the harmonic complex detection algorithm, which we expect will strongly reduce the number of incorrect formants due to incorrectly or not extracted harmonic complexes. Although our results at present are preliminary, we think that these initial results indicate that perceptual features thought to be important for humans in speech processing can also be used to build robust vowel detection systems.

References

[1] Lippmann, R. P. (1997) *Speech recognition by machines and humans.* Speech Communication 22, pp. 1-15.

[2] McClelland, J. L., & Elman, J. L. (1986). *The TRACE model of speech perception.* Cognitive Psychology, 18(**1**), pp. 1-86.

[3] Norris, D. (1994) Shortlist: a connectionist model of continuous speech recognition. Cognition, 52, pp. 189-234.

[4] Luce, P. A., Goldinger, S. D., & Auer, E. T. (2000). *Phonetic priming, neighborhood activation, and PARSYN.* Perception & Psychophysics, 62(**3**), pp. 615-625.

[5] Vargas, J.,& McLaughlin, S. (2008). *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.

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

[7] Hillenbrand, J. M., & Houde, R.A. (2003). *A narrow band pattern-matching model of vowel perception.* Journal of the Acoustical Society of America 113(2), pp. 1044 – 1055.

[8] Wet, F., Weber, K., Boves, L., Cranen, B., Bengio, S., and Bourlard, H. (2004). "*Evaluation of formant-like features on an automatic vowel classification task*," Journal of the Acoustical Society of America 116, 1781–1791.

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

[10] Cooke, M., Green, P., Josifovski, L. & Vizinho, A. (2001) *Robust automatic speech recognition with missing and unreliable acoustic data.* Speech Communication 34, pp 267-285.

[11] O'Shaughnessy, D., (2008). Invited paper: *Automatic speech recognition: History, methods and challenges.* Pattern Recognition 41 (10), 2965–2979.

[12] Hillenbrand, Getty, Clark & Wheeler (1995). *Acoustic characteristics of American English vowels.* Journal of the Acoustical Society of America, 97, pp. 3099-3111.

[13] Witten, I.H., & Frank, E. (2005) *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco.