Robust harmonic complex estimation in noise

J.D. Krijnders, M.E. Niessen and T. Andringa

{j.d.krijnders,m.e.niessen,t.andringa}@ai.rug.nl Artificial Intelligence, *University* of Groningen

Introduction

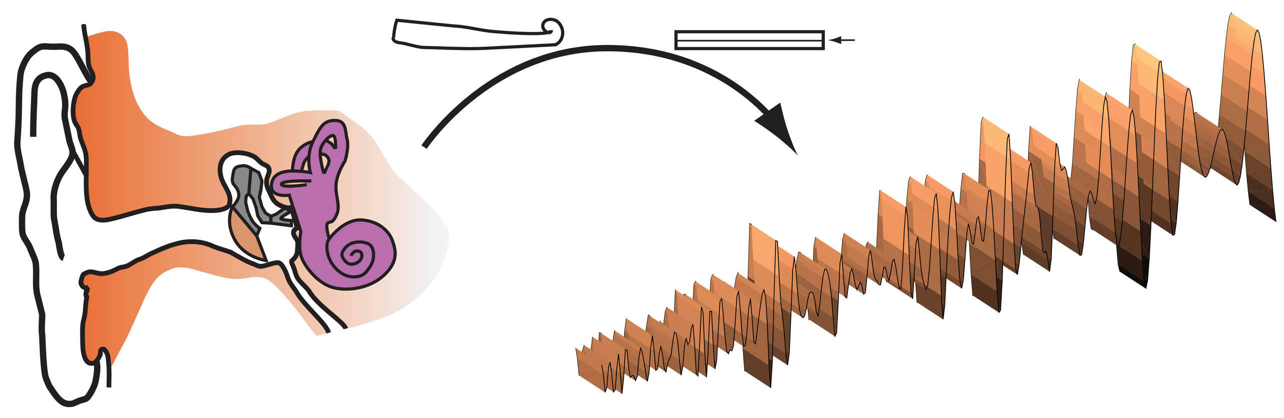
Automatic speech recognition(ASR) systems work only in limited application ranges, because of noise, reverberations and speech variability. This poster addresses the noise problem. Two solution strategies exist:

- Train the recognizer to ignore the noise
- Separate the signal from the noise

Here we focus on separating the signal from the noise. We do this based on fundamental frequency of the voiced parts of speech.

Cochleogram

The proposed method works in the time-frequency plane; the transformation is performed by a model of the human cochlea. The output of this model is the amplitude of the basilar membrane.



The basilar membrane output is leaky-integrated and downsampled. To compress the dynamic range of the energy spectrum, the energy is scaled to a decibel scale. This time-frequency-energy representation is called a cochleogram.

$$E\left(n,t_A\right) = E\left(n,t_A-dt_A\right)e^{-dt_A/\tau_n} + A\left(n,t_A\right) \quad (1)$$

$$E_{dB}\left(n,t\right) = 10log_{10}\left(E_A(n,t)\right) \quad (2)$$

Correlogram

In severe noise conditions the signal begins to disappear in the noise, while it is still clearly audible. One way to extract the fundamental frequency(f_0) is with a running correlogram[?]: a autocorrelate in every channel, at every timestep. For harmonic signals (like much of speech) the channels that correspond to the f_0 have a autocorrelate maximum at the τ corresponding to the f_0.

$$R\left(n,t_A,\tau_{corr}\right) = R\left(n,t_A-dt_A,\tau_{corr}\right)e^{-dt_A/\tau_n} + A\left(n,t_A\right)A\left(n,t_A-dt_A\right) \quad (3)$$

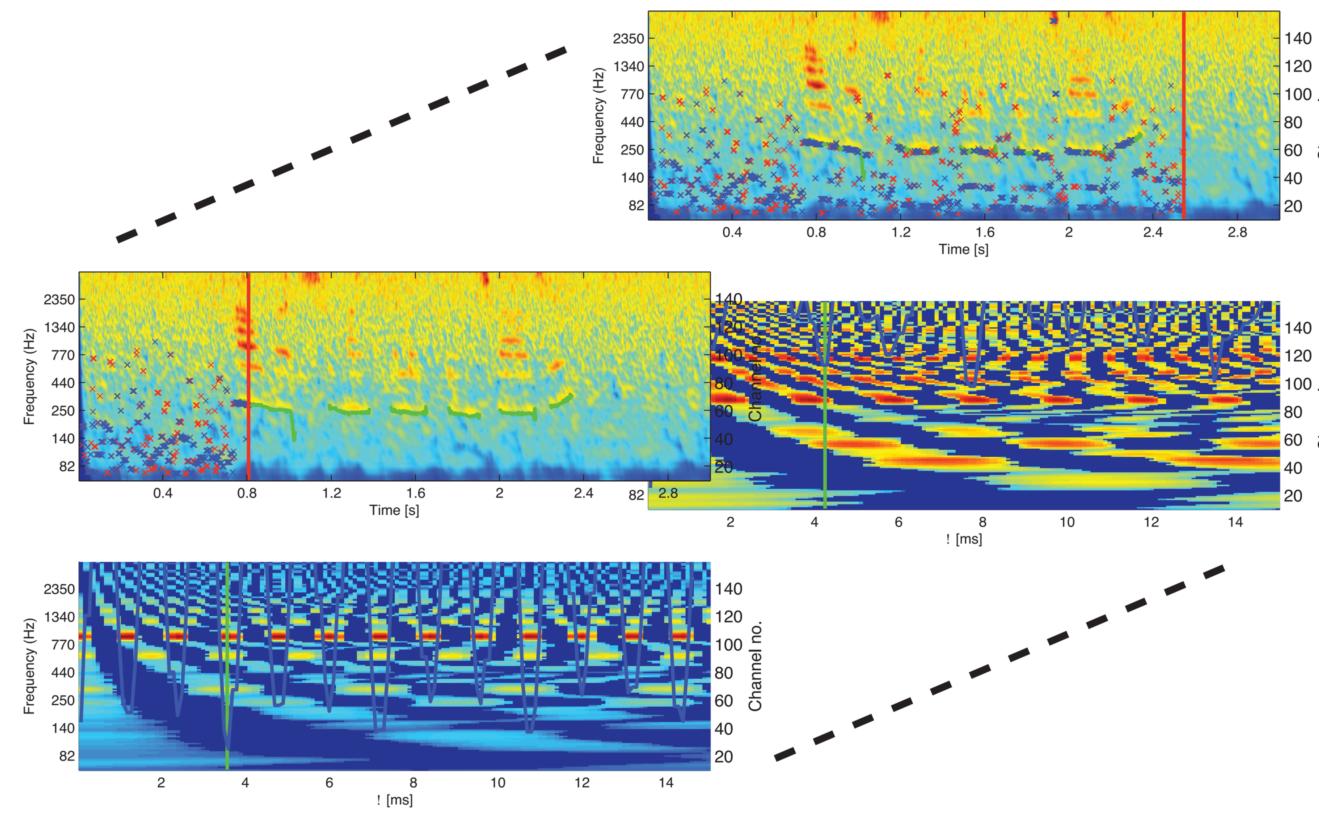


Figure 1: Cochleogram and correlogram with and without signal

Performance measures

The sum of all correlates at a certain time(eqn. **??**) peaks for τ's corresponding to the f_0 and frequencies octaves above and below. (Blue line in figure **??**)

$$m\left(t_A,\tau_{corr}\right) = \sum_{n=1}^{N} R\left(n,t_A,\tau_{corr}\right) \quad (4)$$

The relative height of the peak(eqn. **??**) is a measure of how good defined the pitch is.

$$c = m(peak) - (m(valleybefore) + m(valleyafter))/2 \quad (5)$$

Blue dots in figure **??**

Tracking

The peaks from the measures are connected in subsequent frames allowing a mismatch up to 0.2ms. Due to noise these tracks are generally broken into subtracks, these are connected by dilution, checking for connectedness and subsequent eroding. The confidence in a complete track then becomes:

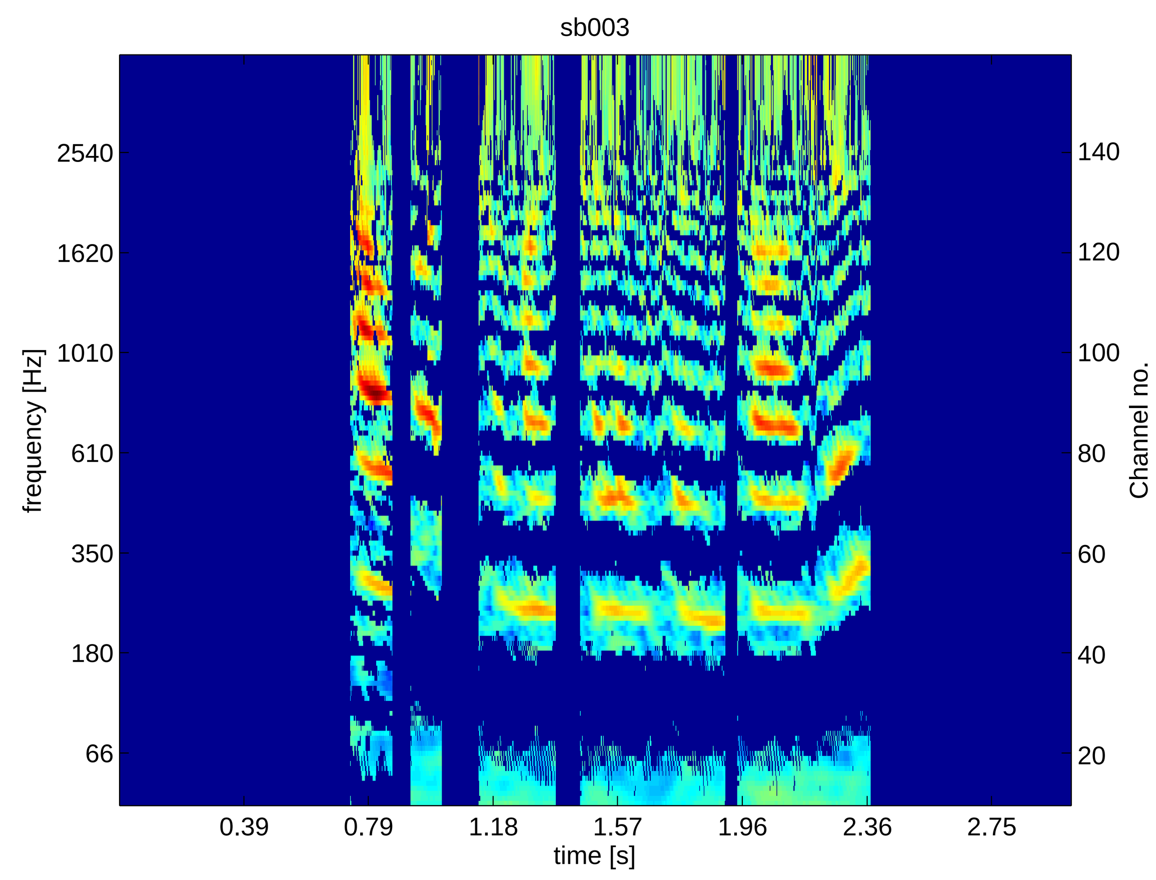
$$c_{track} = \frac{\sum_{track} m(peak) - (m(valleybefore) + m(valleyafter))/2}{\sqrt{L_{track}}} \quad (6)$$

Division by the root of the length is necessary not to favour long tracks too much.

Selection

A threshold for the confidence of a track was trained on a subset of the dataset[**?**]. Tracks with a higher confidence were selected.

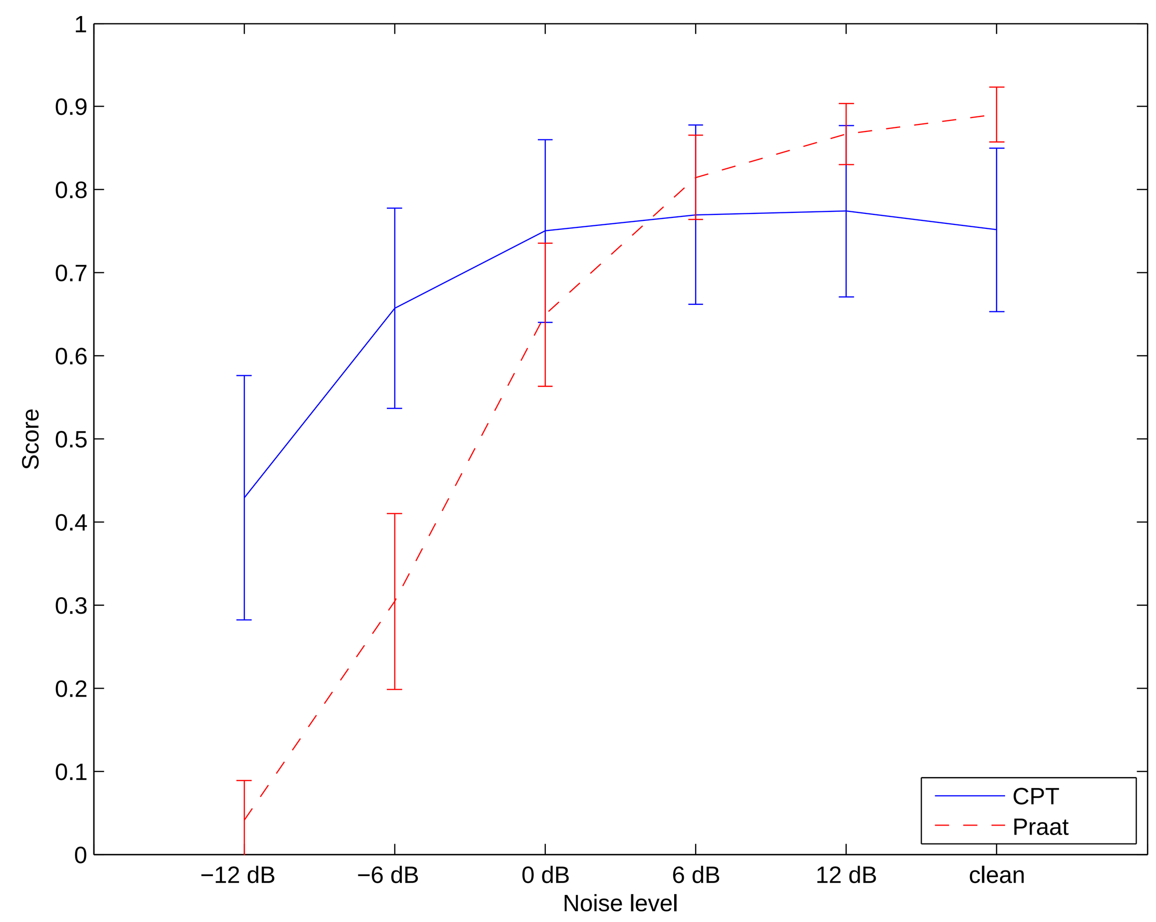
Masks for further (ASR) processing were made by selecting those parts of the cochleogram that showed a positive correlation in the correlogram based on the pitch (see below)



Results

The score is calculated for all 600 signals (2 speakers, 50 sentences, 5 SNRs and the clean signal) as the percentage of time where a pitch is correctly detected, minus the percentage of time where a pitch is incorrectly detected. A chance process scores zero on this measure. The score is compared to that of praat[**?**] on the same dataset.

- Successful selection of harmonic parts of speech
- Higher score for severe noise situations



Further research

- Feedback from the next processing stage can improve performance
- Apply machine-learning techniques to include more features and improve performance
- Include multi-pitch discrimination
- Model non-harmonic speech to increase scores in clean situation
- Other pitch hypothesis sources to increase speed

[1] P.C. Bagshaw, S. Hiller, and M.A. Jack. Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching. In *EUROSPEECH'93*, pages 1003–1006, 1993.

[2] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings*, 17:97–110, 1993.

[3] M. Slaney and RF Lyon. A perceptual pitch detector. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 357–360, 1990.