19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

ROBUST HARMONIC COMPLEX ESTIMATION IN NOISE

PACS: 43.66.Hg

Krijnders, Dirkjan; Niessen, Maria; Andringa, Tjeerd
University of Groningen, Grote Kruisstraat 2/1, 9712 TS, Groningen, the Netherlands; j.d.krijnders@ai.rug.nl; m.e.niessen@ai.rug.nl; t.c.andringa@ai.rug.nl

ABSTRACT
We present a new approach to robust pitch tracking of speech in noise. We transform the audio signal to the time-frequency domain using a transmission-line model of the human ear. The model output is leaky-integrated resulting in an energy measure, called a cochleogram. To select the speech parts of the cochleogram a correlogram, based on a running autocorrelation per channel, is used. In the correlogram possible pitch hypotheses are tracked through time and for every track a salience measure is calculated. The salience determines the importance of that track, which makes it possible to select tracks that belong to the signal while ignoring those of noise. The algorithm is shown to be more robust to the addition of pink noise than standard autocorrelation-based methods. An important property of the grouping is the selection of parts of the time-frequency plane that are highly likely to belong to the same source, which is impossible with conventional autocorrelation-based methods.

INTRODUCTION
To broaden the application range for automatic speech recognition (ASR) it is vital that ASR-systems operate in circumstances where users expect it to work. In general these are the circumstances where human speech recognition is reliable, that is, in almost all situations and environments. Speech in such circumstances is seldom free of noise or reverberations. This poses serious problems for modern speech recognizers that perform optimally in narrow application domains and have no means to deal with the increased variability due to unknown environmental influences. We propose a method to separate quasi-periodic speech from noise, or, at least, indicate where an ASR system might try to recognize speech.

Two completely different approaches might solve this problem: either train the recognizer to ignore the noise or separate the target speech from the noise. The first choice is the standard approach of ASR systems that have originally been developed for clean speech [1]. This approach does not normally include sound source separation, but treats noise as a perturbation that has to be discounted. This would require all types of noise the system may be confronted with to be present in the training set, while maintaining sufficient specificity for detection.

The second approach is more difficult because no specific information about the signal is available, which entails that the separation needs to be based on general speech properties [2]. The method proposed in this article takes this approach, because it focuses on the use of robust information that can be used select regions of the time-frequency plane that are dominated by a quasi-periodic speech-like event. We call our method correlogram-based pitch tracking (CPT). It separates the harmonic parts of speech from noise by creating a mask that marks parts of the spectrum as target (speech) and the rest as background.

Many algorithms that determine the pitch of a signal exist [3], mainly focus on accuracy, for example, YIN [4]. The algorithm presented here is similar to the correlogram method proposed by Slaney and Lyon [5], but the auditory model is significantly simplified. We add a pitch track selection mechanism and a way to extract individual harmonics. The generation of this selection is useful as input of an ASR system.

METHODS
The proposed method works in the time-frequency plane; the transformation is performed by a model of the human cochlea [6]. This model includes outer and middle ear transfer functions.

The basilar membrane is modeled as a transmission-line. We did not model the haircells, but instead calculated the energy directly from the squared basilar membrane displacement by leaky-integration (eq. 1). The leaky-integration time constant τ is different in every channel and was set to 2 times the characteristic period of the channel.

The energy matrix was down-sampled to 200 Hz (eq. 2), which is safe because leaky-integration acts as a low pass filter. To compress the dynamic range of the energy spectrum, the energy is scaled to a decibel scale (eq. 2). This time-frequency-energy representation is called a cochleogram.

$$E_A(n,t_A) = E_A(n,t_A - dt_A)e^{-dt_A/\tau_n} + A(n,t_A) \quad \text{(eq. 1)}$$

$$E_{dB}(n,t) = 10\log_{10}(E_A(n,t)) \quad \text{(eq. 2)}$$

Where *n* is the channel number, which corresponds to frequency via the place-frequency relation of the greenwood map [7], *t* is discrete time, *dt* is the discrete time step, *A* is the amplitude of the basilar membrane and *E* is energy. The subscript *A* denotes a representation at the sample frequency of the transmission-line (22400 Hz). An example of the output *E* can be seen in figure 1.

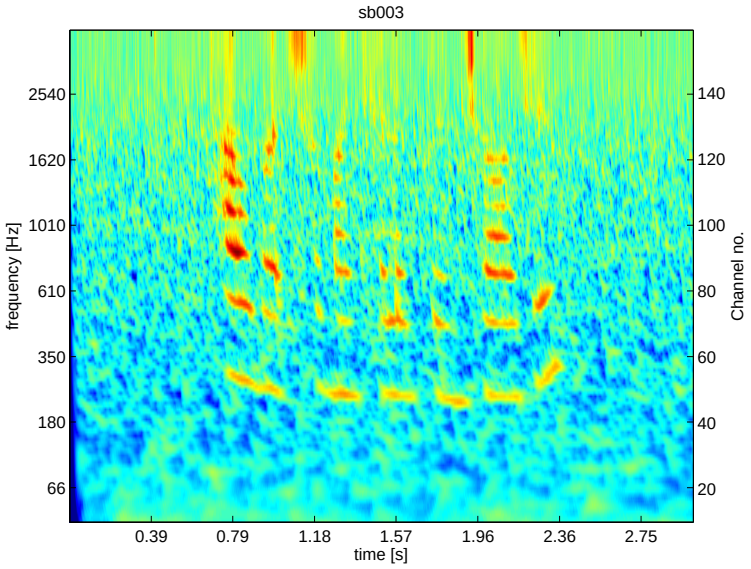

figure 1.-Cochleogram of a woman saying “How much are my telephone charges”

Apart from the cochleogram calculation, the amplitude of the basilar membrane is auto-correlated within each channel and leaky-integrated over time:

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

An example of such a running auto-correlate can be seen in figure 2. *R* is called a correlogram. A minus sign is chosen for the correlate as is standard procedure; a plus sign may however increase performance [8]. For harmonic signals, all channels that have a center frequency at integer multiples of the fundamental frequency show high correlation values for the periodicity corresponding to that fundamental frequency. Summing the correlates over all *N* channels (eq. 4) gives a measure of how well the signal correlates with itself for that delay. This is the blue line in the lower picture of figure 2. It clearly shows peaks at the correct periodicity (3.7 ms) and octaves above and below.

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

To form pitch tracks, the maxima in *m* are tracked through time, allowing jumps up to 0.2 ms. For every track, the peak height relative to the neighbouring valleys is recorded.

The next step selects pitch tracks based on their cumulative peak height, divided by the square root of their length. We divided by a length measure, because the accumulation favours longer tracks disproportionally (see eqs. 5 and 6). A threshold for determining which tracks belong to the target was set, optimizing for the sum of the scores (see results section).

$$c_{track} = \sum_{track} m(peak) - \left(m(valleybefore) + m(valleyafter)\right)/2 \text{ (eq.5)}$$

$$c_{track} = c_{track} / \sqrt{L_{track}} \text{ (eq.6)}$$

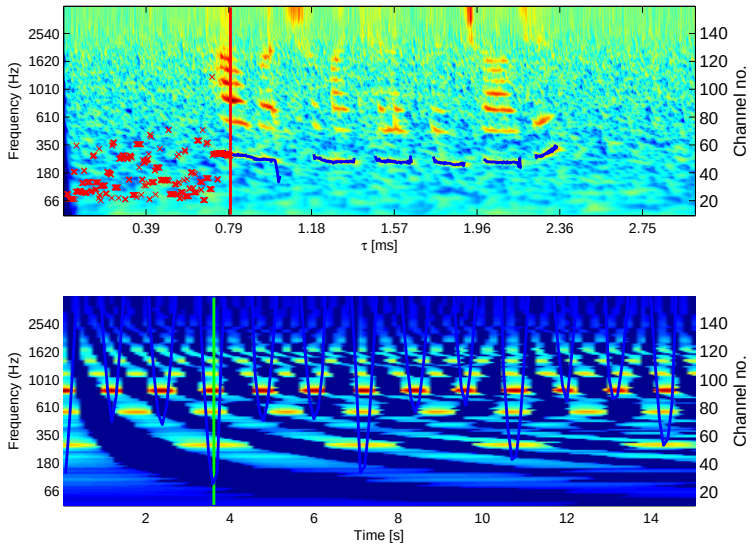

figure 2.-Correlogram instance at t = 0.79 s. The red crosses show the relative peak maximum for each time. The blue lines in the upper picture are the larynxographs(see below).

Due to noise, tracks are often broken into sub tracks. Once the candidate tracks are selected other tracks are checked for nearness (8-connectedness [9]) at the begin and end points and added to the candidate if feasible. The resulting tracks are the final choice for the pitch tracks.

All cochleogram regions with a positive correlation for the selected pitch tracks form a mask. The complete mask can be seen in figure 3. This mask is applied to the cochleogram to filter out the noise contributions. figure 5 shows the signal with noise removed. This cleaned cochleogram can then be transformed into the input of standard ASR techniques by some form of envelope estimate, based on connected harmonic energies.

The signals for testing our algorithm are taken from the Bagshaw dataset [10] and pink noise was artificially added at levels -12, -6, 0, 6, 12 dB signal to noise ratio. figure 1 gives an example of 0 dB SNR. The Bagshaw dataset consists of 50 sentences spoken by both a male and a female speaker. For all recordings a larynxograph, a recording of the actual vibration of the larynx, is available, which is used as a ground truth defining the time intervals where pitch is present.

RESULTS

The performance of our algorithm is compared to the performance of PRAAT[11] on the same dataset. Praat uses a standard autocorrelation algorithm. First a comparison is made whether or not the algorithms are capable of correctly detecting pitch in noisy signals.

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 results for Praat and our algorithm are shown in figure 4.

Due to the nature of the algorithm most errors are octave errors; the correct pitch is usually the second hypothesis. These errors can be avoided by taking some context into account. For less noisy conditions other methods exist that can easily raise the peak close to unity.

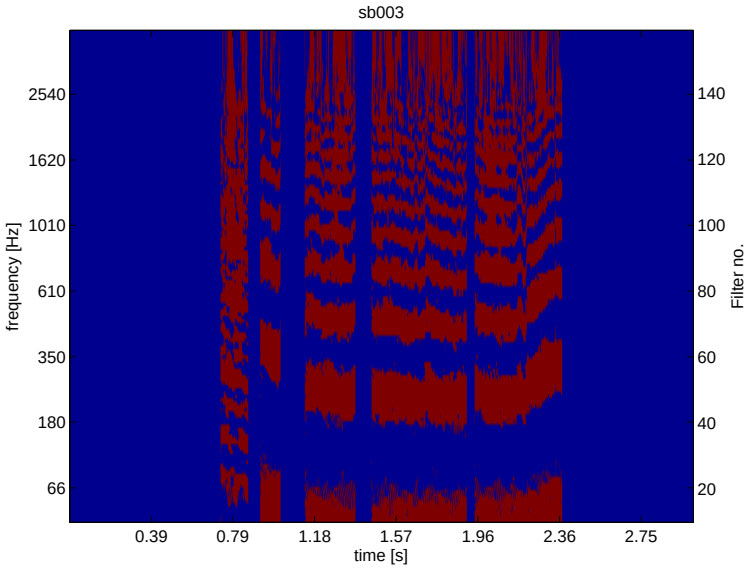

figure 3.-Mask derived the correlogram

CONCLUSIONS AND DISCUSSION

We have shown that we are able to select pitch tracks from signals in very severe noise conditions. Scores for the clean signals are lower than existing methods due to the absence of the modeling of the non-harmonic parts of the speech, although this modeling is possible [5].

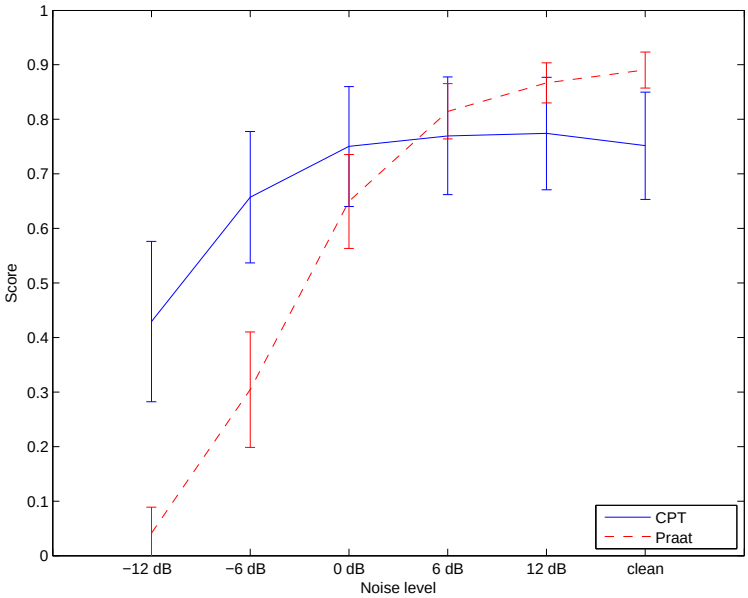

figure 4.-Average frames correct score with standard deviation

Although reasonable results are shown based on a single threshold, other measures and their combinations could lead to better results. Machine learning techniques for finding these combinations and their optimal threshold could further increase performance. This could also include multi-pitch discrimination.

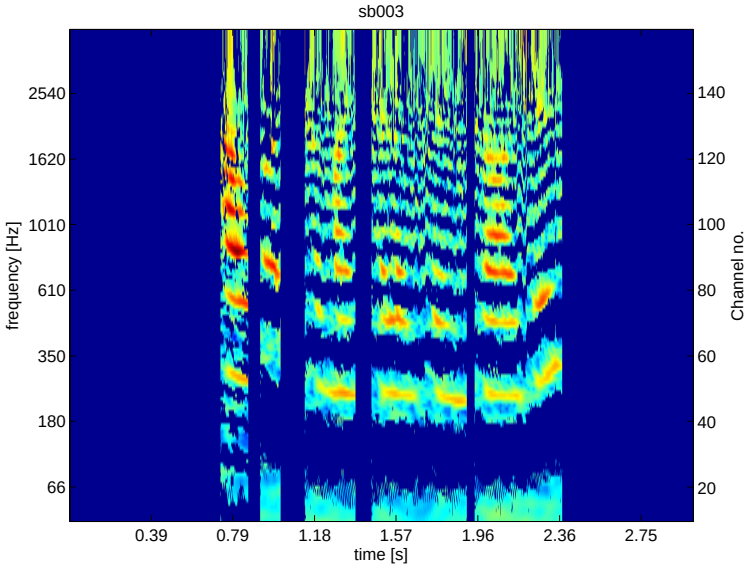

figure 5.-Masked cochleogram, the end result of the algorithm

Performance is a major issue when calculating autocorrelates. It could significantly be increased when an assumption is made about the height of the pitch, for example limit it between 100 Hz and 400 Hz. Also methods that work directly on the cochleogram could give hypotheses for the pitch.

We have shown a pure data-driven selection algorithm, but feedback from a, still to be developed ASR system, might be used to select different pitch tracks. This combination may be necessary to reach good recognition results in arbitrary circumstances. For example, the second best hypothesis of the track selection algorithm may be included in the ASR input giving it more freedom to optimize its output. However, this is not possible with traditional ASR methods that do not allow for this type of flexible interaction between signal processing and pattern recognition.

Acknowledgement
The support of the Dutch Science Foundation NWO under grant 634.000.432 (CASSANDRA) within the ToKeN2000 program is gratefully acknowledged.

References:[1] J.C. Junqua, J.P. Haton: Robustness in automatic speech recognition: Kluwer Academic Publishers (1996)
[2] Qifeng Zhu, Markus Iseli, Xiaodong Cui, Abeer Alwan: Noise Robust Feature Extraction for ASR Using the Aurora 2 Database: Eurospeech'01 (2001)
[3] W.J. Hess: Pitch determination of speech signals: Springer, New York (1993)
[4] A. de Cheveigné, H. Kawahara: YIN, a fundamental frequency estimator for speech and music: Journal of the Acoustical Society of America **111 no 4** (2002) 1917-1930
[5] M. Slaney, R.A. Lyon: A perceptual pitch detector: International Conference on Acoustics, Speech, and Signal Processing 1990, 357-360
[6] H. Duifhuis, H. Hoogstraten, S. Netten, R. Diependaal, W. Bialek: Modelling the cochlear partition with coupled Van der Pol oscilators: Cochlear mechanics: Structure, Function and Models, J. Wilson, D. Kemp (Ed.), Plenum, New York (1985) 395-404
[7] D.D. Greenwood: Critical bandwidth and the frequency coordinates of the basilar membrane: Journal of the Acoustical Society of America **33** (1961) 1344-1356.
[8] T.C. Andringa: Continuity Preserving Signal Processing; PhD. thesis University of Groningen (2002)
[9] F. van der Heijden: Image Based Measurement Systems: John Wiley and Sons, New York (1994)
[10] P.C. Bagshaw, S. Hiller, M.A. Jack: Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching: EUROSPEECH'93 (1993), 1003-1006.
[11] P. Boersma: PRAAT, a system for doing phonetics by computer: Glot International **5, no 9/10** (2001) 341-345