

Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score

Kris Hermus, *Member, IEEE*, Hugo Van hamme, *Member, IEEE*, and Sufian Irhimeh

Abstract—We present a new algorithm for the estimation of the voicing cut-off frequency (VCO), i.e. the frequency that separates the harmonic low-frequency part from the aperiodic high-frequency part in voiced speech. The VCO is estimated as the frequency for which the sum of the harmonicity scores of all pitch harmonics below that frequency, is maximised. The algorithm is combined with a powerful dynamic programming approach to track the VCO estimates over time. Remarkably accurate and smooth VCO contours are obtained, despite the simplicity of the algorithm. Applications include a.o. (sinusoidal) speech modelling, coding and synthesis, as well as harmonic speech analysis for e.g. automatic speech recognition.

Index Terms—Speech analysis, speech coding, speech synthesis, speech processing.

I. INTRODUCTION

In various domains of speech processing, the spectrum of a voiced speech frame is analysed in terms of its degree of harmonicity. In harmonic-plus-noise modelling (HNM), the speech signal is decomposed in a harmonic series of sinusoids and a noise signal. In the feature extraction module of automatic speech recognisers, the decomposition of speech into a harmonic and a stochastic part provides useful information for the further disambiguation of basic speech units or phones [1].

The observation that for most speech frames the harmonic structure is most pronounced in the lower part of the spectrum has motivated researchers to split the spectrum in two distinct parts: a low-frequency harmonic part and an aperiodic high-frequency part [2]. The separation between both parts occurs at the so-called *Voicing Cut-Off frequency* (VCO). The location of the VCO is phone specific, and within-phone variations occur due to pitch movements, co-articulation effects, etc. In practice, a speech spectral slice almost never shows two clearly distinct parts. Rather, it is very common to find adjacent harmonic-like and noise-like bands in the mid-range frequencies. Moreover, random pitch fluctuations (jitter) may cause the high frequency harmonics to appear inharmonic due to the frequency modulation within the analysis window. These observations make the estimation of the VCO a difficult problem using spectral analysis.

Fortunately, the adoption of one single split-point in the spectrum – the VCO – has proven to be a reasonable and effective approximation of reality. A well-known example is harmonic-plus-noise modelling of speech that produces very natural sounding speech, in which the well-known buzziness and “lack of fullness” of pure sinusoidal speech is eliminated.

II. VCO ESTIMATION ALGORITHMS

Numerous VCO estimation algorithms exist, which are broadly based on one or more of the following approaches:

Analysis-by-Synthesis (AbS) For these algorithms, first the voicing degree is expressed in terms of the goodness-of-fit of the analysed speech to a dedicated speech model (e.g. a harmonic sinusoidal model). This voicing degree is then mapped onto the VCO. AbS methods usually do not account for the *distribution* of the harmonic and noise energy along the frequency axis. Hence, it is not unlikely that the VCO is placed in the middle of a series of well-identified harmonics, or in the middle of a clearly aperiodic spectral region. An algorithm that belongs to this class can be found in [3].

Spectral domain methods These methods inspect individual harmonics of the speech signal, and put the split-point between the voiced and unvoiced parts at the frequency where harmonicity seems to disappear. The latter is always based on some heuristic criterion since the definition of harmonicity is a subjective matter. In Stylianou’s algorithm [4] the speech spectrum is classified as either voiced or unvoiced (binary decision) at every pitch harmonic, based on the peakiness of a high-resolution FFT and on the deviation of that peak from its expected location (the latter is a means to deal with the effects of jitter). The resulting series of ones (voiced harmonic) and zeros (unvoiced harmonic) is smoothed by a three point median filter. The spectrum is considered to be voiced up to the first pitch harmonic classified as unvoiced. A drawback of this approach is that (1) the binary classification of the harmonic candidates is very sensitive to the empirical thresholds in the definition of peakiness, and (2) the determination of the VCO based on a three point median filtering of the 1s and 0s cannot account for the phenomenon of alternating periodic and aperiodic spectral regions in natural speech.

Time domain methods In this case, the VCO estimate is based on a measure of the periodicity of the (filtered) time signal. An interesting algorithm was published by Kim et al. [5]. For every candidate value f of the VCO, a periodicity score for the low frequency band ($0-f$ Hz) and a non-periodicity score for the high frequency band ($f-f_s/2$ Hz) are calculated (with f_s being the sampling frequency). Both scores are based on the time autocorrelation function. The VCO is defined as the frequency for which the sum of both scores – called the *combined subband periodicity score* (CSPS) – reaches a maximum. A drawback of the algorithm is that the objective function often does not have a clear maximum. Another algorithm that falls into this class can be found in [1].

III. PROPOSED TECHNIQUE

The algorithm that we propose combines ideas from the algorithm of Stylianou [4] (further referred to as the Harmonic

Inspection (HI) method) and from the CSPS function of Kim et al. [5]. With our algorithm we try to overcome the above described limitations of the HI and the CSPS methods. Our algorithm operates on a normalised power spectrum from which the degree of harmonicity of each candidate harmonic is derived, followed by the determination of the maximum of a cumulative harmonicity score. In other words, we are looking for a division of the spectrum such that the low frequency part is maximally harmonic, according to a local and global harmonicity score. We assume that an accurate pitch contour of the speech signal is available. We used the Praat software [6] for automatic pitch labelling, but many other excellent algorithms have been described in literature.

Unvoiced speech segments are assigned a VCO of 0 Hz. To the voiced speech portions, a framing with a variable frame length (see below) and with a fixed frame shift (typically 2 to 5 ms) is applied. For each voiced speech frame with a pitch frequency of p Hz, we estimate the number of voiced pitch harmonics h , and obtain the VCO as the product $h p$.

Spectral estimation Let $s(k)$, $k = 1 \dots N$ be a speech frame of 2 pitch periods length, with corresponding discrete Fourier transform (DFT) $S(k)$, $k = 1 \dots N$ based on a rectangular windowing. Without loss of generality and to simplify the derivation, we assume that N is odd, and only consider the first $(N+1)/2$ DFT coefficients, i.e. the “positive” frequencies of the DFT. It is clear that the odd lines ($k = 1, 3 \dots (N+1)/2$) and the even lines ($k = 2, 4 \dots (N-1)/2$) of $S(k)$ contain pitch periodic (or harmonic) components and pitch aperiodic (or noise) components of $s(k)$, respectively¹. Apart from the DC component, which can be assumed zero without loss of generality, we have $K = (N-1)/4$ pitch harmonic lines from which the degree of harmonicity has to be examined.

Peakiness We switch to the power spectrum $P(k) = |S(k)|^2$, and define the degree of harmonicity $H(j)$ (or peakiness) of the power spectrum for the j^{th} ($1 \leq j \leq K$) candidate harmonic as follows:

$$H(j) = \frac{P(2j+1)}{P(2j+1) + \left[\frac{P(2j) + P(2j+2)}{2} \right]} \text{ for } j = 1 \dots K$$

In words, the peakiness $H(j)$ is based on the degree to which the power spectrum at the candidate harmonic exceeds the average value of its neighbouring non-harmonic power spectral lines². Observe that $0 \leq H(j) \leq 1$.

Nonlinear transformation We apply a nonlinear transformation to the peakiness score to obtain a harmonicity score $\bar{H}(j)$ in the range $[-1, 1]$. This non-linear transformation implements a soft thresholding, mapping harmonic spectral lines close to 1 and random lines to zero or lower. After empirical optimisation, we found that the following expression

¹In fact, if the speech frame contains a noise part, the harmonic part will be perturbed. However, simulations have shown that compensating for this bias is not necessary for our purposes.

²Other measures are possible, e.g. based on the maximal neighbouring power spectral value.

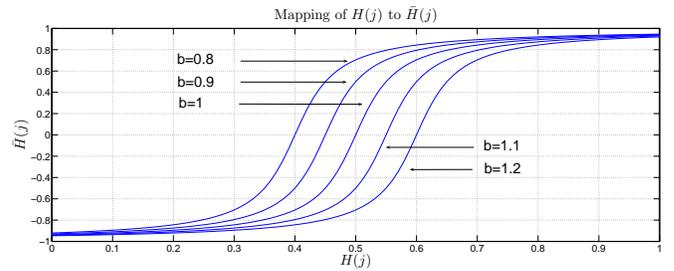


Fig. 1. Mapping functions of $H(j)$ to $\bar{H}(j)$ for five different values of the control parameter b .

for $\bar{H}(j)$ performs very well:

$$\bar{H}(j) = \frac{2}{\pi} \arctan(10(2H(j) - b))$$

with b a control parameter that is used to set the sensitivity of the harmonicity score. Decreasing the value of b , decreases the threshold for harmonicity and leads to higher VCO estimates. The optimal value will mostly depend on the user preference and/or on the application (e.g. sometimes we want the VCO to be the frequency below which the spectrum is undoubtedly voiced, sometimes we rather like it to be the frequency above which no harmonic-like spectral parts exist).

From experiments, we found that $b = 1.2$ is a good compromise. Note that the value of b is fixed for all speech frames. The nonlinear mapping functions for a few common values of b are given in figure 1.

Cumulative harmonicity score We now calculate the cumulative harmonicity score $C_h(j)$ that is associated to the first j pitch harmonics

$$C_h(j) = \sum_{i=1 \dots j} \bar{H}(i), \text{ for } j = 1 \dots K$$

For a given number of harmonics $1 < j < K$, $C_h(j)$ is a measure of the degree of harmonicity for the spectral bandwidth from 0 to $j p$ Hz. In practice, the cumulative score $C_h(j)$ will be positive for small values of j , and continue to increase as long as harmonic candidates are added that exceed the threshold of harmonicity. Aperiodic pitch harmonics will contribute negatively such that $C_h(j)$ will be globally decreasing as soon as the spectrum becomes aperiodic. The location of the global maximum of $C_h(j)$ is assumed to be the split point between the low-frequency harmonic and high-frequency aperiodic part of the speech spectrum.

Maximisation The VCO is now defined as the number of harmonics j for which $C_h(j)$ maximised, multiplied by the frame pitch p

$$\begin{aligned} \text{VCO} &= \left(\arg \max_j C_h(j) \right) p \\ &= h p \end{aligned}$$

Our algorithm is illustrated in figure 2 for the phoneme /e/ (as in *tail*). The high-resolution FFT is *not* used in the algorithm, but is depicted here for illustration purposes only.

Note: There is a caveat w.r.t. the DFT-based spectral estimation if one pitch period $T (= f_s/p)$ does not contain an integer

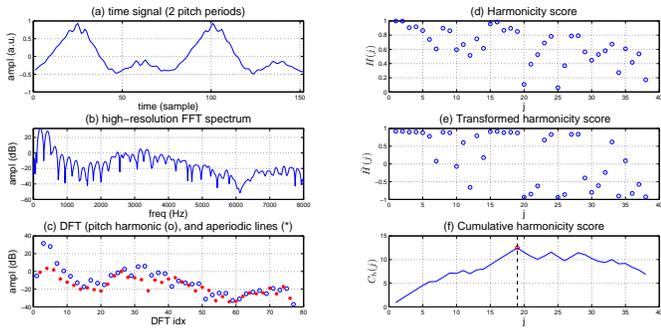


Fig. 2. Illustration of the VCO estimation algorithm: (a) time signal, (b) high-resolution FFT, (c) DFT with pitch harmonics (o) and aperiodic lines (*), (d) harmonicity score $H(j)$, (e) transformed harmonicity score $\tilde{H}(j)$ with $b = 1.2$, (f) cumulative harmonicity score $C_h(j)$ with its maximum.

number of samples. If we take $N = \text{round}(2T) = 2T + \alpha$ with $-0.5 < \alpha \leq 0.5$, the DFT will actually evaluate the spectrum at the frequencies $(l/(2T+\alpha))f_s \approx l(p/2)(1-\alpha/N)$ with $l = -(N-1)/2 \dots (N-1)/2$ (assume N is odd without loss of generality) instead of at the frequencies $l(p/2)$. The error $l(p/2)(\alpha/N)$ grows linearly with l to reach its maximum value of approx. $\alpha(p/4)$ for the highest harmonic, which is significant, but acceptable. It can be shown that for a perfectly periodic signal, the value of $H(j)$ for the highest harmonic can decrease from 1 ($\alpha = 0$) to 0.93 ($\alpha = 0.5$). However, an accurate estimation of the spectral amplitudes at the *exact* frequencies $l(p/2)$ is possible by solving the set of equations $\mathbf{E}\mathbf{x} = \mathbf{s}$ in least squares (LS) sense, with \mathbf{E} an $N \times M$ matrix ($M = 1 + 2 \lfloor \text{floor}(f_s/p) \rfloor$) with $e^{j\pi(k-1)(l-\frac{M+1}{2})p/f_s}$ on row k and column l , \mathbf{x} a column vector containing the estimates of the spectral amplitudes, and \mathbf{s} a column vector with $s(k)$, $k = 1 \dots N$. It is also possible to calculate a zero-padded FFT spectrum and to retain the spectral values closest to $l(p/2)$ Hz. We will come back to this in section V.

IV. TIME SMOOTHING

Algorithms for the determination of the VCO usually result in VCO values that exhibit moderate, or sometimes large, frame-to-frame variations. The main reason for these time fluctuations is the lack of robustness of the estimation algorithms to the ill-posed problems they have to solve. In order to reduce these fluctuations, most algorithms include a dedicated smoothing technique. Examples are a.o. averaging of adjacent VCO estimates, or heuristic decision-based smoothing. While these approaches are often effective to reduce major frame-to-frame variations, they are not suited to model the smooth variations in the VCO contour that are found in transitional segments of natural speech (e.g. co-articulation effects).

Here we propose a totally different approach to obtain a smooth - and accurate - VCO contour. Instead of independently extracting the maxima of the individual cumulative harmonicity score functions, we look for a smooth path through a series of these score functions, based on dynamic programming (DP). Mathematically, for a speech utterance of L frames, we find the smoothed series of number of voiced pitch harmonics $\tilde{h}_1 \dots \tilde{h}_L$ using a constrained Viterbi

algorithm in which the following total score is maximised:

$$\tilde{h}_i \leftarrow \arg \max_{\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_{i+F}} \sum_{j=1}^{i+F} \left[t_{\tilde{h}_{j-1}, \tilde{h}_j} + C_h(\tilde{h}_j) \right]$$

with $t_{k,l}$ the score for a transition from k harmonics in the current frame to l harmonics in the next frame, and F the number of frames of the *look-ahead*. In other words, \tilde{h}_i at frame i is found on the basis of information up to frame $i + F$. For F sufficiently large, the Viterbi path at frame i tends not to depend on data beyond frame $i + F$, but to avoid abrupt changes, search paths not passing through \tilde{h}_i are pruned. We assume that the speech is embedded in silence, such that $h_0 = 0$. If not, another appropriate initialisation should be made.

Since the optimal path is to a large extent dependent on the values of the transition scores $t_{k,l}$, a well-considered choice has to be made. First, smoothness can be controlled by excluding transitions that induce large frame-to-frame changes in the number of harmonics (note that this is dependent on the frame update rate, and on the pitch). Second, the smoothness can also be influenced by the relative difference between the different transition scores. From our experiments, we found that excellent results are obtained if only frame-to-frame changes of maximal 2 to 4 harmonics are allowed (for a 3ms frame shift), and if all other transitions are excluded. Good values for the allowed transitions $t_{k,k\pm l}$ are e.g. 10, 8, 7 (for $k < 25$ and $l = 0, 1, 2$), and 10, 8, 7, 7, 6 (for $k \geq 25$ and $l = 0, 1, 2, 3, 4$). Allowing more transitions for high k is a means to account for the fact that in the case of a lower pitch, the number of voiced harmonics should be allowed to faster change over time. Note that excluding unlikely transitions also significantly reduces the search space and limits the computational load.

The look-ahead strategy that is embedded in the dynamic programming formulation, leads to an algorithmic delay, but experiments indicated that a look-ahead of 20 ms already yields very accurate results in most cases. However, many coding applications exist for which the algorithmic delay is of no importance (e.g. storage of compressed speech). A well-known example application that we target in our work is the sinusoidal coding of the speech database of a corpus-based text-to-speech (TTS) system [4]. In this case a larger look-ahead can be used since it will increase the accuracy of both the pitch estimation and the VCO estimation.

V. EVALUATION

The VCO algorithm was extensively tested on a speech database containing sentences of Dutch female speech, as well as on male and female English spoken utterances. The original sampling frequency f_s was 16 kHz, but also downsampled speech at 8 kHz was used for testing. An illustration is given in figure 3, from which the graceful adaptation of the VCO contour to the time-varying signal characteristics can be observed. The solid line is the result for a DP search with 15 ms look-ahead, whereas the dashed line is obtained with a full search DP (= one global Viterbi optimisation on all frames in the utterance). From the plot it is clear that using a small look-ahead hardly compromises the result.

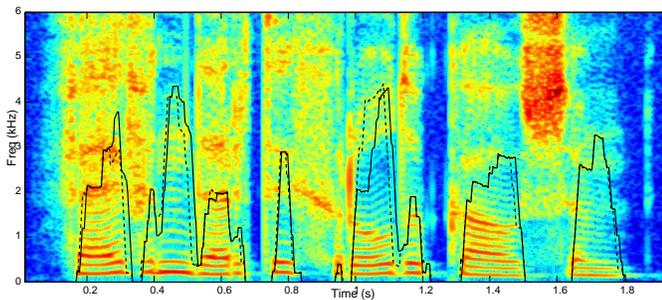


Fig. 3. Smoothed VCO contour with a DFT-based spectral estimation for an utterance of the Dutch words *vijfendertig rode kousen*, with 15 ms look-ahead (solid) and full search (dashed).

A. Simulations

From our tests we have found that very accurate and smooth VCO contours can be obtained, and that we have very good control over the threshold of harmonicity - i.e. the degree of sensitivity to peakiness of the spectrum - via the parameter b . We found that the *low-complexity DFT method works almost equally well as the higher complexity LS method*, even though there is evidence that the higher harmonics are not very well estimated with the DFT method. We believe that the lack of accuracy of the DFT method is to a large extent compensated by the redundancy caused by a high frame rate. However, for low sampling frequencies there is a tendency to underestimate the VCO. This can be solved by using a smaller value for the control parameter b (e.g. 1.1 instead of 1.2). In general, the DFT method will need a slightly larger threshold b to obtain a VCO contour with similar properties as the LS-based method.

For comparison, we also implemented the HI [4] and CSPS [5] methods. The HI algorithm was combined with two consecutive median filterings of order 5. This is in line with the author's suggestion, and proved to generate smooth VCO contours. We found that the smoothing method that was proposed for the CSPS method was not able to provide sufficient smoothness. We therefore combined the algorithm with an adapted version of our Trellis-based tracking technique.

From our simulations, we found that our algorithm provides more accurate - and definitely more consistent - results than the CSPS method, for which sometimes large frame-to-frame fluctuations in the estimated VCO were observed. These errors cannot always be solved by smoothing. Our algorithm also performs better than the HI technique, mostly due to our powerful tracking technique that cannot be combined with the HI method. Moreover, we are convinced that our algorithm is easier to control, since only one important tuning parameter b has to be optimised. Finally, we mention the low computational complexity of our algorithm, with a DFT per frame being the dominant factor in the computations.

B. Experiment

A subjective evaluation was set up with 10 speech files. For each file, the VCO contour was estimated with our method, as well as with the HI and CSPS methods. Each of the 30 VCO contours was plotted on a narrowband spectrogram from the speech file it was estimated from. We presented the

spectrograms to nine subjects³ active in speech research and asked them to classify each VCO contour on a scale from 1 (very inaccurate) to 5 (very accurate)⁴. All subjects were briefed about the kind of errors that typically occur in VCO estimation, and about their seriousness.

The mean scores - averaged over the 9 participants in the test and over the 10 files in the test - are 3.71 for our algorithm, 2.70 for the HI method, and 2.46 for the CSPS method. To check whether these means are statistically different, we applied the non parametric Wilcoxon signed-rank test for paired observations. We obtain a P-value of 0.002 under the null hypothesis that the median of the difference between the matched scores of our method and the HI method, is zero. The P-value is the probability of obtaining a test statistic as extreme or more extreme than the one observed under the assumed null hypothesis. Likewise, we find a P-value of 0.002 for a comparison between our method to the CSPS. These results provide strong evidence that the new algorithm is superior to the reference methods.

VI. CONCLUSIONS

The definition of peakiness of a spectrum is a subjective matter, such that the determination of the VCO is a complex problem. In our VCO estimation algorithm, the peakiness of every harmonic candidate is examined, and the VCO is found by finding the maximum of a cumulative harmonicity score function. The degree of peakiness that is required for a spectral band to be classified as a true pitch harmonic can be set via a control parameter, which makes it possible to adapt the behaviour of the VCO estimator to the user's preferences. We introduced a dynamic programming approach that has proven to provide excellent tracking of the VCO, while the algorithmic delay is kept minimal. Subjective experiment have shown that our algorithm is superior to both the HI and the CSPS algorithms.

REFERENCES

- [1] L. Gu and K. Rose, "Split-band perceptual cepstral coefficients as acoustic features for speech recognition," in *Proc. European Conference on Speech Communication and Technology*, Sep. 2001, pp. 583-586.
- [2] J. Makhoul, R. Viswanathan, R. Schwartz, and A. Huggins, "A mixed-source model for speech compression and synthesis," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Apr. 1978, pp. 163-166.
- [3] R. McAulay and T. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995, pp. 121-173.
- [4] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 1, pp. 21-29, Jan. 2001.
- [5] E.-K. Kim, W.-J. Han, and Y.-H. Oh, "A new band-splitting method for two-band speech model," *IEEE Signal Process. Lett.*, vol. 8, no. 12, pp. 317-320, Dec. 2001.
- [6] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 4.4.23) [computer program]," retrieved May 30, 2006, from <http://www.praat.org/2006>.

³The developers of the new algorithm did not participate in the experiment.

⁴The randomly mixed VCO contours were presented at once, and repeated views of the contours were allowed.