ESTIMATION OF THE VOICING CUT-OFF FREQUENCY CONTOUR OF NATURAL SPEECH BASED ON HARMONIC AND APERIODIC ENERGIES

Kris Hermus[¶][*], *Laurent Girin*[†], *Hugo Van hamme*[¶], *and Sufian Irhimeh*[‡]

[¶]Dept. ESAT, Katholieke Universiteit Leuven, Belgium
{kris.hermus,hugo.vanhamme}@esat.kuleuven.be

[†]GIPSA-lab, INPG, Grenoble, France
laurent.girin@gipsa-lab.inpg.fr

[‡]Nuance Communications Belgium
sufian.irhimeh@nuance.com

ABSTRACT

We present a new algorithm for the automatic estimation of the voicing cut-off frequency (VCO), i.e., the frequency that separates the periodic low-frequency part from the aperiodic high-frequency part in voiced segments of natural speech. Starting from the power spectrum of a two pitch period speech frame, we define the VCO to be located at the frequency for which the sum of the periodic and aperiodic energy in the spectral band below and above that frequency respectively, is maximised. By formulating the problem in terms of a score function we are able to apply a dynamic programming based smoothing technique. Remarkably smooth and accurate VCO contours were obtained, despite the simplicity of the proposed algorithm. In a formal evaluation the algorithm compares favourably to two existing VCO estimation techniques.

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

1. INTRODUCTION

In various domains of speech signal processing the spectrum of a speech frame is analysed in terms of its degree of harmonicity. 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 harmonic low-frequency part, and an aperiodic high-frequency part [1]. The transition between both parts occurs at the *voicing cut-off frequency* (VCO).

An important application of this concept is found in e.g., harmonic-plus-noise modelling (HNM) and coding, in which the speech signal is represented by the combination of a series of harmonically related sinusoids and a synthetic noise signal. HNM is successfully used for high-quality corpus-based speech synthesis, and it is suited for high-quality speech modification as well [2]. In automatic speech recognition systems, the decomposition of speech into a harmonic and a stochastic part has proven to provide useful information for the disambiguation of phones [3].

The location of the VCO can vary a lot from one section of a speech signal to another. It is determined mainly by turbulent flow in the vocal tract (e.g., fricative, plosive or aspiration noise), and also by cycle-to-cycle irregular variations in the shape of the glottal excitation signal (shimmer), as well as by the perturbation of the fundamental frequency (jitter). Moreover, even a regular evolution of the speech signal within the analysis frame (e.g., F0 melody of speech) can have an influence on spectral characterisation. 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. This makes the definition of a hard transition point in the spectrum (the VCO) an ill-posed problem, and its estimation is a heuristic process. Fortunately, such a two-band model has proven to be a reasonable and effective approximation of reality, as illustrated by the very naturally sounding HNM speech [2], despite the low number of parameters involved in this representation.

Several algorithms for the estimation of the VCO exist. Spectral *analysis-by-synthesis* methods like, e.g., [4, 5, 6], are based on the goodness-of-fit of the short-time speech spectrum to a harmonic sinusoidal representation. Those methods usually do not account for the distribution of the harmonic and noise energy along the frequency axis. It is therefore not unlikely that with this strategy the VCO is placed in the middle of a series of well-identified harmonics, or in the middle of a clearly aperiodic spectral region.

This problem is circumvented by methods that inspect individual harmonics of the speech signal, and put the HN split-point at the frequency where harmonicity seems to disappear [1]. In [2], Stylianou classifies the speech spectrum at every pitch harmonic as either voiced or unvoiced (binary decision), based on the peakiness of a high-resolution FFT spectrum and on the deviation of that peak from its expected location (the latter is a means to deal with the effects of e.g., jitter). The resulting series of ones (voiced harmonic) and zeros (unvoiced harmonic) are further smoothed by a 3rd order median filter. The spectrum is then considered to be voiced up to the first pitch harmonic classified as unvoiced. The heuristic nature of the procedure leads to the adoption of parameters for which the optimal values are set empirically.

Alternatively, time-domain methods for VCO estimation exist, which are generally based on a measure of the periodicity of the (filtered) time signal, using autocorrelation functions [7, 3]. In [8], 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 the sampling frequency). Both are based on the time autocorrelation function. The VCO is then defined as the frequency for which the sum of both scores reaches a maximum.

[*] Kris Hermus is a Postdoctoral Research Fellow of the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Flanders), project ‘SAMSSoNS’.

In the next section, we propose a new algorithm for VCO estimation. This algorithm is innovative but it also combines some ideas from literature mentioned above [2, 8] and from our recent work [9]. It operates on a normalised power spectrum from which cumulative periodic and aperiodic subband energy scores are combined and maximised. Also, it is important to mention that for all algorithms, excessive frame-to-frame variations of the estimated VCO are almost unavoidable. These variations can be significantly reduced by applying an appropriate smoothing technique, and we describe in section 3 a smoothing procedure to be associated with our VCO estimation process. The overall process is evaluated in section 4.

2. PROPOSED TECHNIQUE

We assume that an accurate pitch contour of the speech signal is available. The pitch extraction tool from the Praat software [10] proved to perform excellently for our purpose, but many other excellent algorithms have been proposed. Unvoiced speech segments are assigned a VCO of 0 Hz. To the voiced speech portions a variable-length framing is applied, with a fixed frame shift in the range 2-5 ms. For each voiced speech frame with a pitch frequency of p Hz, we now estimate the number of harmonics h, and the VCO is given as $h.p$. The successive steps of the algorithm are as follows :

Spectral estimation The VCO estimation algorithm that we propose is based on the observation that the Fourier transform of a speech frame that contains an integer number of pitch periods, is an excellent starting point for splitting that speech frame into a harmonic and an aperiodic part. This principle is used, e.g., in the time domain harmonic scaling (TDHS) method of Malah [11], and in the harmonic/noise separation technique of Jackson and Shadle [12] who considered signal frames of four pitch-periods.

Let here $s(k)$, $k = 1 \dots N$ be a speech frame of two pitch periods length[1], with corresponding discrete Fourier transform (DFT) $S(k)$, $k = 1 \dots N$ based on a rectangular windowing :

$$S(k) = \sum_{i=1}^{N} s(i) \exp^{-j2\pi(k-1)\frac{(i-1)}{N}}, \; k = 1 \dots N$$

Without loss of generality and to simplify the derivation, we assume that N is odd, and we 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 $s(k)$, respectively. If the speech frame contains a noise part, it will be spread over all DFT coefficients, and the harmonic part will be perturbed. If we assume that the noise part has a relatively smooth unvoiced spectrum, then – on average – approximately half of its energy will be contained in the harmonic part.

Apart from the DC component, which can be assumed zero without loss of generality, we have $K = (N-1)/4$ pitch harmonic spectral lines and K spectral lines that contribute to the aperiodic part. We now try to find the number h $(1 \le h \le K)$ of pitch harmonics that are well resolved and that clearly exceed the neighbouring aperiodic spectral components. In other words, we are looking for a division of the spectrum into a low-frequency part that is maximally harmonic, and a high-frequency part that is maximally aperiodic.

Normalisation Before proceeding we switch to the power spectrum $P(k) = |S(k)|^2$, and perform a spectral equalisation (normalisation) step, inspired by the observation that the estimated degree of harmonicity of a spectrum should be independent of the spectral envelope. We define the normalised power spectrum as :

$$P_n(k) = \begin{cases} 1 & \text{for } k = 1 \\ \frac{(P(k)+P(k+2))/2}{(P(k)+P(k+2))/2+P(k+1)} & \text{for } k = 2, 4, \dots (N-1)/2 \\ \frac{P(k)}{(P(k-1)+P(k+1))/2+P(k)} & \text{for } k = 3, 5, \dots (N+1)/2 \end{cases}$$

Observe that two neighbouring values of $P_n(k)$, with $k > 1$ always add up to 1, and that $\sum_{k=1}^{(N+1)/2} P_n(k) \approx K$.

Cumulative energy We now calculate the cumulative periodic energy $E_h(j)$ that is associated with the first j pitch harmonics. In other words, $E_h(j)$ is a measure of the degree of harmonicity for the spectral bandwidth from 0 to $j.p$ Hz:

$$E_h(j) = \sum_{i=3,5,\dots 1+2j} \max\left(0, P_n(i) - P_n(i-1)\right), \text{for } j = 1 \dots K$$

The noise subtraction in the above calculation is similar to the Wiener filtering process applied in the sinusoidal framework [13].

We also calculate the cumulative aperiodic energy $E_a(j)$ that is associated with the spectral part *starting from* the j^{th} pitch harmonic *and up*. In other words, $E_a(j)$ is a measure of the degree of aperiodicity of the spectral bandwidth from $j.p$ Hz to f_s:

$$E_a(j) = 2 \sum_{i=2j,2j+2,\dots(N-1)/2} P_n(i), \text{for } j = 1 \dots K$$

Combining cumulative energies The VCO is now defined as the number of harmonics j for which the combination of $E_h(j)$ and $E_a(j)$ is maximised, multiplied by the frame pitch p

$$\text{VCO} = \left(\arg\max_j E_h(j) + bE_a(j) \right) . p$$

in which b is a parameter that controls the behaviour of the VCO estimator. The higher b, the lower the average estimate of the VCO will be. The optimal value will mostly depend on the application (e.g., sometimes we want the VCO to be the frequency below which the spectrum is mostly voiced, sometimes we rather like it to be the frequency above which little harmonic-like spectral part is identified). From experiments, we found that the value of b is typically in the range 0.2 - 0.6. Note that the value of b is fixed for all speech frames. An illustration of VCO estimation from cumulative energies is given in section 4.

Possible refinements : In practice, if one pitch period T $(= f_s/p)$ does not contain an integer number of samples, the DFT will not evaluate the spectrum at *exactly* the pitch frequencies and its harmonics. In [9] we have shown that the error that we make is acceptable, even at the highest harmonics. An accurate estimation of the spectral amplitudes at the *exact* frequencies is always possible by performing a least squares (LS) spectral estimation instead of a DFT-based one. Also, this can be combined with a refinement of the pitch measure. This is what we do in the calculation of the spectrograms in section 4, but the method has been shown to work very well without those refinements.

3. TIME SMOOTHING

As mentioned in the introduction, any frame-based algorithm for the determination of the VCO can result in significant 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. Note that even when the VCO of an individual

[1] Fewer pitch periods means less frequency modulation within the analysis window, which leads to a more reliable spectral estimation.

speech frame is manually annotated by different human experts, significant differences can be observed. If the speech context (e.g., by providing the spectrogram) is known, a more consistent VCO estimate can be found using a smoothing procedure.

Most algorithms that are found in literature apply some local smoothing technique, e.g., averaging over some neighbouring frames, or heuristic decision-based smoothing when e.g., the difference in measured VCO between adjacent frames exceeds some threshold. However, the effectiveness of these approaches in producing correctly smoothed VCO contours is mostly limited.

Here we use a totally different approach to obtain a smooth and accurate VCO contour that we presented in [9]. Instead of independently extracting the maxima of the individual combined cumulative energy functions, we calculate a *constrained* (smooth) path through a series of these score functions, using dynamic programming (DP). For a speech utterance containing L frames, we find the smoothed series of values $\tilde{h}_1 \ldots \tilde{h}_L$ using a constrained Viterbi algorithm in which the following total score is maximised :

$$\tilde{h}_i \longleftarrow \underset{\tilde{h}_1,\tilde{h}_2\ldots\tilde{h}_{i+F}}{\arg\max} \sum_{j=1}^{i+F} \left[t_{\tilde{h}_{j-1},\tilde{h}_j} \;+\; (E_h(\tilde{h}_j) + b\,E_a(\tilde{h}_j))\right]$$

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 $h_0 = 0$. Otherwise, another appropriate initialisation should be made.

Since both the smoothness and the location of 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 a frame-to-frame change in the number of harmonics of more than e.g., 5 (note that this is dependent on the frame update rate). Second, the smoothness can also be influenced by the relative difference between the different transition scores.

Based on simulations with a 3ms frame shift, we found that excellent results are obtained if only frame-to-frame changes of maximally 2 (high pitched speech) to 5 (low pitched speech) harmonics are allowed, and if all other transitions are excluded. If the gender of the speaker is unknown, a good compromise is to set $t_{k,k\pm l} = 10, 8, 7, 7, 6, 6$ ($\forall k$ and $l = 0, 1, 2, 3, 4, 5$). Allowing larger transitions is a means to account for the fact that in the case of a lower pitch and/or fast speaking rate, the number of voiced harmonics should be allowed to change faster 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 DP process leads to an algorithmic delay. For off-line applications (e.g., storage of compressed speech in corpus-based text-to-speech (TTS) system [2]) the algorithmic delay is of no importance and can be user-specified. For on-line applications, experiments indicated that a reasonable look-ahead of 20 ms is sufficient in most cases for obtaining reliable smooth VCO trajectories.

4. EVALUATION

4.1. Illustration

The VCO estimation algorithm is illustrated in figure 1 for the phoneme /e/ (as in t*ai*l). The high resolution FFT is not used in the algorithm but it is depicted here for illustration purposes only. It can be seen that the harmonic structure of the spectrum in figure 1.b is visible in the lower part, whereas the upper part is much more noisy. Consequently, harmonic and noise bins are much more tied together in the upper part of figure 1.c. As illustrated in figure1.d, the cumulative energy criterion is effective in capturing the boundary between those regions.

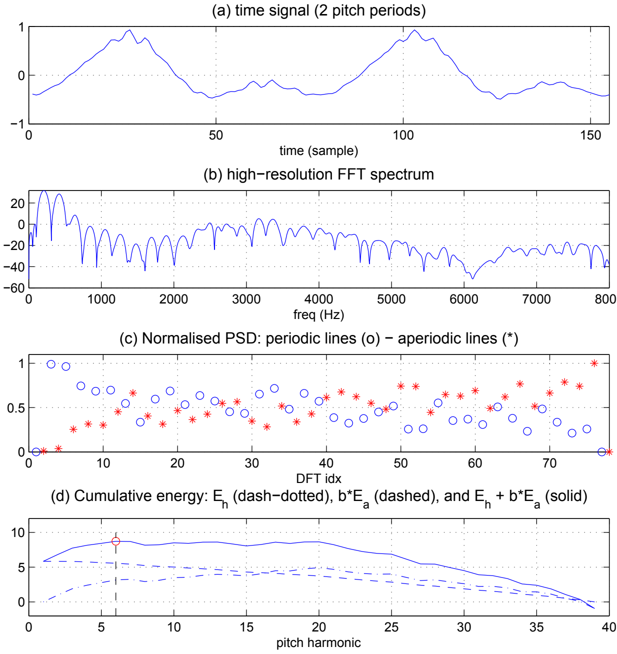

Fig. 1. Illustration of the VCO estimation algorithm : (a) time signal, (b) high-resolution FFT, (c) normalised power spectrum with periodic lines and aperiodic lines, (d) $E_h(j)$ (dash-dotted), $b.E_a(j)$ (dashed), and their sum (solid) with $b = 0.4$.

The parameters of the VCO algorithm and the Viterbi tracking have been optimised based on a training database containing sentences of Dutch female speech, as well as on male and female English spoken utterances. An illustration is given in figure 2 (top part), from which the smooth adaptation of the VCO contour to the time-varying signal characteristics can be observed.

From our tests we have found that very accurate and smooth VCO contours can be obtained, and that the estimated VCO contours gracefully adapt to changes in the value of control parameter b.

4.2. Experiment

For the evaluation of the new algorithm we have set up a formal experiment in which 11 subjects have participated (the authors not included). We randomly selected 16 speech files (8 male, 8 female) from a large database of English speech (the test files were *not* used for the tuning of the algorithm). For each file, we estimated the VCO contour with our new method, and with Kim et al.'s method (CSPS) and Stylianou's method (HI), implemented as described in [2] and [8], respectively. Following the author's suggestion, the HI algorithm was combined with two consecutive median filterings of order 5. This proved to generate smooth VCO contours. We found that the smoothing method 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.

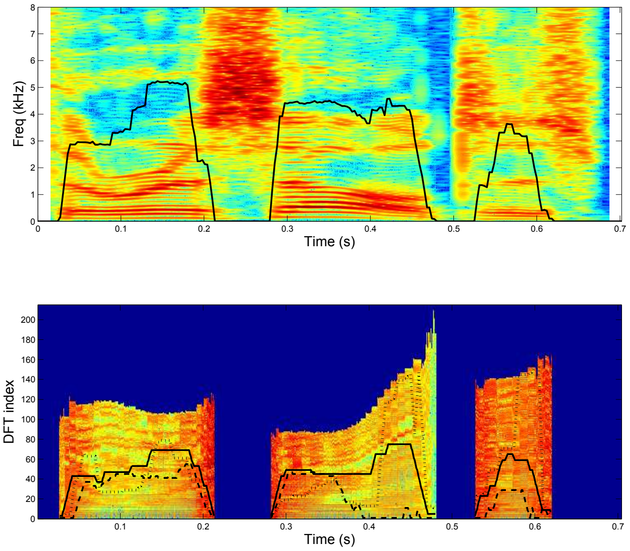

Fig. 2. Example of VCO annotation for an utterance of the words *the result is*. Top : narrowband spectrogram with VCO contour obtained with the new method. Bottom: dedicated spectrogram with VCO curve obtained by our method (solid), the HI method (dotted), and the CSPS method (dashed). Only voiced parts were analysed.

From each test file we created a spectrogram based on speech frames of two pitch periods length. In order to prevent the test persons from being misled by the speech formants, we removed the formant information by an inverse linear prediction (LP) filtering. The unusual resulting spectrograms guarantee that the harmonic character of the speech (if present) is visible as much as possible. An illustration is given in figure 2 (bottom part).

To each spectrogram we added the VCO contours estimated by the three methods (the linestyles were randomly mixed for each test file). The subjects were briefly informed about the VCO concept, and were asked to rate the quality of each VCO contour on a scale from 1 to 5 (1 for much below average, 5 for much above average). Repeated views as well as rescoring were allowed.

The mean scores – averaged over the 11 subjects and over the 16 files – are 3.91 for our algorithm, 2.03 for the HI method, and 1.97 for the CSPS method. The score differences are statistically significant: a Wilcoxon signed-rank test for paired observations gives a P-value of 0.00043 for our method vs. the HI method, and a P-value of 0.00044 for our method vs. the CSPS. These results clearly show that the new algorithm provides more accurate results than the reference methods. Moreover, our algorithm performs definitely more consistently over a large pitch range. The HI estimator has a tendency to underestimate the VCO, and the CSPS seems to be very sensitive to changes in the voiced/unvoiced ratio, which leads to large frame-to-frame fluctuations in the estimated VCO (even when combined with an adapted Trellis-based tracking technique).

Besides, the results indicate that our method performs slightly better for female than for male speech. The HI method performs better for male speech than for female speech, and the CSPS estimator produces quite disappointing results for male speech.

5. CONCLUSIONS

The estimation of the VCO is a heuristic and ill-posed problem. We have shown that the spectrum of a two pitch-period speech segment, combined with the sum of the cumulative periodic and aperiodic energies of the lower and higher frequency band, is a simple but excellent basis for the separation of this spectrum into a periodic and a aperiodic part. A control parameter in the summation enables the user to set the degree of peakiness that is required for a spectral band to be classified as harmonic.

We also described a dynamic programming approach that has proven to provide excellent VCO contours, while the minimal algorithmic delay is low and the overall complexity remains reasonable.

Even though the algorithm yields very accurate results in most cases, there is still room for improvement, especially for low pitched voices. Probable extensions are e.g., spectral estimation based on prefiltered speech, and the use of 3 pitch period frames (more robust spectral estimation but with reduced time resolution). These extensions will further increase the robustness, but at the expense of a higher computational load.

6. REFERENCES

[1] J. Makhoul, R. Viswanathan, R. Schwartz, and A.W.F. Huggins, “A mixed-source model for speech compression and synthesis,” in *Proc. ICASSP*, Apr. 1978, pp. 163–166.

[2] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. on SAP*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[3] L. Gu and K. Rose, “Split-band perceptual cepstral coefficients as acoustic features for speech recognition,” in *Proc. EUROSPEECH*, Sept. 2001, pp. 583–586.

[4] R.J. McAulay and T.F. Quatieri, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Eds., pp. 121–173. Elsevier, 1995.

[5] C.A. Rødbro, J. Jensen, and R. Heusdens, “Adaptive time-segmentation of speech for packet loss channels,” Tech. Rep., Delft University of Technology, 2003.

[6] I. Atkinson, S. Yeldener, and A.M. Kondoz, “High quality split-band LPC vocoder operating at low bit rates,” in *Proc. ICASSP*, Apr. 1997, pp. 1559–1562.

[7] C. Laflamme, R. Salami, R. Matmti, and J.-P. Adoul, “Harmonic-stochastic excitation (HSX) speech coding below 4kbits/s,” in *Proc. ICASSP*, May 1996, pp. 204–207.

[8] E.-K. Kim, W.-J. Han, and Y.-H. Oh, “A new band-splitting method for two-band speech model,” *IEEE SP Letters*, vol. 8, no. 12, pp. 317–320, Dec. 2001.

[9] K. Hermus, H. Van hamme, and S. Irhimeh, “Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score,” *IEEE SP Letters*, vol. 14, no. 11, pp. 820–823, Nov. 2007.

[10] 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/.

[11] D. Malah, “Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals,” *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 121–133, Apr. 1979.

[12] P.J.B. Jackson and C.H. Shadle, “Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech,” *IEEE Trans. on SAP*, vol. 9, no. 7, pp. 713–726, Oct. 2001.

[13] J. Jensen and J.H.L. Hansen, “Speech enhancement using a constrained iterative sinusoidal model,” *IEEE Trans. on SAP*, vol. 9, no. 7, pp. 731–740, Oct. 2001.