A MIXED SPEECH F_0 ESTIMATION ALGORITHM

Alain de Cheveigné

Laboratoire de Linguistique Formelle,
CNRS-Université Paris 7, Paris, France.

ABSTRACT

This mixed speech fundamental frequency (f_0) estimation algorithm is an extension of the classical AMDF (Average Magnitude Difference Function) algorithm for one voice. An exhaustive search of the parameter space of two cascaded time-domain comb filters yields an estimation of the periods of the component voices. The algorithm, which is computationally expensive but easily parallelizable, was tested on a database of continuous male and female speech. Segments of voiced speech, selected according to a "good periodicity" criterion to ensure that the reference single-voice f_0 algorithm would not fail (this criterion rejected 25% of voiced speech frames), were paired and summed to simulate mixed speech. The search range of the algorithm was limited to a 3 octave range, and search was performed frame-by-frame without continuity constraints. The resulting estimates were compared to those of the reference algorithm and found to be within 3 % of target values for 90 % of all frames.

Keywords: speech, fundamental frequency, pitch extraction, mixed speech separation, noise reduction, cocktail-party effect.

INTRODUCTION

Adverse conditions of noise, reverberation or interfering speech greatly affect the performance of speech processing devices. Interfering speech is particularly troublesome, because it shares the spectro-temporal characteristics of the target speech, and because it violates the assumptions of estimation methods such as LPC. Human listeners, however, appear to maintain intelligibility by using a variety of sources of information, both high-level (familiarity with the lexicon or language, understanding of the meaning, etc.) and low-level (binaural or fundamental frequency disparities) (Cherry 1953).

As demonstrated with steady-state synthetic vowels, a difference between the fundamental frequencies of concurrent vowels makes them easier to identify (Assmann and Summerfield 1990; Scheffers 1983). In one experiment, the recognition rate for both vowels correct was about 55% for identical fundamentals, and 75% for a difference of two semitones (Assmann and Summerfield 1990). When the fundamentals are the same, the stimulus sounds like a single vowel, "colored" by the identity of a second vowel. When they differ, the stimulus sounds like two talkers pronouncing different vowels with different pitches.

A number of models or algorithms have been proposed to account for, or reproduce, our ability to separate speech using a difference in fundamental frequency (Parsons 1976; Frazier, Samsam, Braida and Oppenheim 1976; Nagabuchi, Kobayashi and Yamamoto 1979; Scheffers 1983; Kitamori, Harada and Kawarada 1984; Weintraub 1985, 1986; Palmer 1988, 1990; Stubbs and Summerfield 1988, 1990; Assmann and Summerfield 1990; Duda and Lyon 1990; Meddis and Hewitt 1990). Most of these models require at some point that both fundamental frequencies be estimated. This is a difficult task. For a mixture of stationary synthetic vowels differing in fundamental frequency by one semitone, the algorithm of Scheffers correctly estimated one fundamental out of two (within 3% of the value used for synthesis) for 96% of all frames, and both for 24%. For connected digits pronounced concurrently by a male and a female speaker, Weintraub reported period estimates within 5 samples (.31 ms) of target values for 88.8% ("dominant" voice) and 74.3% ("weaker" voice) of the frames for which both channels were voiced. The reliability of f_0 extraction is insufficient for voice separation algorithms to be of practical use.

Most of the many algorithms that have been proposed for the extraction of f_0 from a single isolated voice (Hess 1983) rely on regularities in the time domain (regular occurrence of remarkable points, similarity between periods, etc.) or in the frequency domain (regularly-spaced frequency components, etc.). When the signal consists of two simultaneous voices, the spectral or temporal patterns overlap and the resulting pattern is difficult to interpret. Indeed, in several of the methods mentioned above, fundamental frequency estimation is closely dependant on the speech separation process itself.

This paper describes an algorithm that attempts to estimate the fundamental periods of mixed voiced speech by modelling the mixed speech signal as the sum of two *periodic* signals.

THE ALGORITHM

By definition, a signal S is periodic of period T, if for all t:

$$S(t) = S(t+T)$$

If we feed this signal to a comb-filter defined by its impulse response $\delta(t) - \delta(t+\tau)$, the output is identically zero if the lag τ is equal to the period T, or its multiple. This is the basis of a classical f_0 estimation algorithm known as AMDF (Average Magnitude Difference Function) (Ross, Shaffer, Cohen, Freudberg and Manley 1974). The lag parameter space of a comb-filter is searched for a minimum of the AMDF function:

$$\mathrm{AMDF}(\tau) = \int_W \left| S(t) - S(t+\tau) \right| dt$$

The lag at which the minimum occurs indicates the period.

Likewise, if we feed a signal S that is the sum of two periodic signals of periods T_A and T_B to two cascaded comb filters of impulse response $\delta(t) - \delta(t+\tau_A)$ and $\delta(t) - \delta(t+\tau_B)$, the output is identically 0 if $\tau_A = T_A$ and $\tau_B = T_B$. This is the basis of the mixed speech estimation algorithm described in this paper. The two dimensional lag parameter space of two cascaded comb filters is searched for a global minimum of the Double Difference Function (DDF):

$$\mathrm{DDF}(\tau_A, \tau_B) = \int_W \left| S(t) - S(t+\tau_A) - S(t+\tau_B) + S(t+\tau_A+\tau_B) \right| dt$$

Other minima can occur, corresponding to lags equal to period multiples: the algorithm avoids them by choosing the smallest lags, or by restricting the search range.

In principle the algorithm is *guaranteed* to find the two periods, unless one is a multiple of the other. In practice, real speech might not be sufficiently periodic for the algorithm to succeed.

EVALUATION

The principle of evaluation is to compare the results of the mixed voice algorithm to those obtained separately on the isolated speech by a reference single voice f_0 estimation algorithm.

• **reference algorithm**

This algorithm is a variant of the AMDF method (Hess 1983; Ross, al. 1974). Speech, sampled at 20 kHz, is smoothed by convolution with a 1ms rectangular window. The AMDF is calculated using overlapping 20 ms rectangular windows at 1.5 ms intervals, over a range of lags corresponding to f_0's of 60 to 300 Hz for a male speaker and 100 to 600 Hz for a female speaker. The value for each lag is divided by the mean of values for shorter lags (to eliminate the zero at zero lag and attenuate spurious dips at short lags), and the minimum of this function is taken as the period. This algorithm can inappropriately lock on to a period multiple (subharmonic). To avoid this, a period minimum is further required to be less than 0.9 times the value at 1/2 or 1/3 its lag. No other smoothing or error correction is used. Period values are transformed to a base 2 logarithmic frequency scale expressing octaves relative to 110 Hz.

The algorithm produces as a by-product a value that can be interpreted as a *measure of periodicity*. This is defined as:

$$PM = \log_2\left(\frac{\text{mean(AMDF)}}{\text{AMDF(period)}}\right)$$

This measure is large (2 to 6) during steady state voiced portions and small (close to 0) at transitions and during unvoiced portions. It gives an indication of the reliability of the f_0 estimate produced by the algorithm.

• **database**

Test data, derived from the ATR database (Kuwabara, Sagisaka, Takeda and Abe 1989), consisted of 3 Japanese sentences pronounced according to five different intonation patterns by one male (known as MYI), and one female speaker (known as FST), a total of 30 sentences. Speech was sampled at 20 kHz, 12 bits resolution, and processed by the reference AMDF algorithm. The periodicity measure was then scanned for runs with a value greater than an arbitrary threshold ($PM > 1.4$) for a duration greater than 225ms. The corresponding portions of speech were excised. Nine such 225 ms segments of speech signal were selected for each speaker. They were paired and summed to obtain "mixed speech" (producing 32 male-male tokens, 32 female-female tokens, and 81 male-female tokens).

The motivation for selecting portions with a good periodicity was to ensure that the reference f_0 tracks used for evaluation were reliable. In the raw data, about 75% of all voiced portions (defined conservatively as any portion with $PM > 0.5$ for more than 30 ms) had a periodicity measure above this threshold.

• **mixed voice algorithm**

The mixed voice f_0 estimation algorithm was implemented using exactly the same window size and analysis increment as the reference AMDF algorithm. Search for each lag parameter was limited to a 3 octave range, adjusted (on the basis of the reference f_0 tracks) to exclude lags longer by 10% or more than the longest period of the target component (the search ranges were thus not necessarily the same for both parameters). This search range constraint eliminated subharmonic errors, which the algorithm cannot avoid. As in the case of the reference algorithm, period values were transformed to a base 2 logarithmic frequency scale.

• **results**

Figures 1, 2 and 3 show typical results. The values produced by the mixed voice f_0 algorithm follow very closely the reference f_0 values. Figures 2 and 3 show how the algorithm breaks down when the f_0's are practically equal. It fails because, when the periods are equal, a single comb filter is sufficient to cancel both voices: the other lag parameter is unconstrained. Similar effects occur when one f_0 is at the octave of the other.

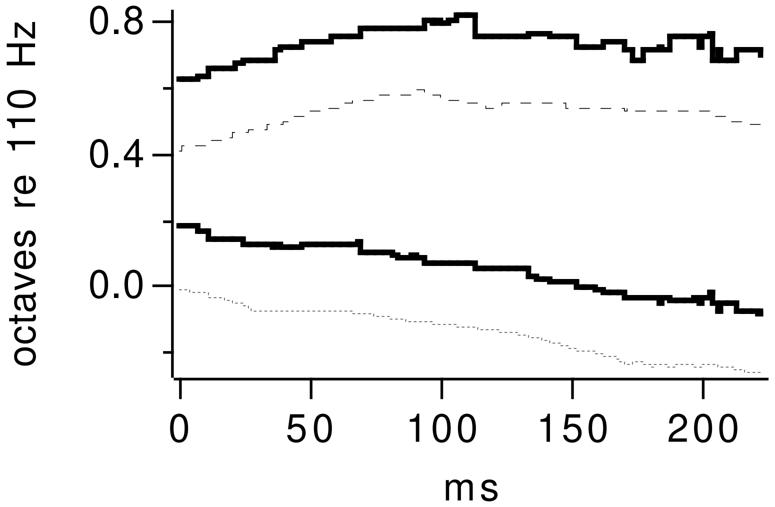

Fig. 1 F_0 estimates for both voices as a function of time. Continuous lines: f_0 values produced by the mixed speech algorithm. Dotted lines (offset by -0.2 octaves for clarity): reference f_0 values obtained from speech before mixing. Male speaker, tokens a0 and a1.

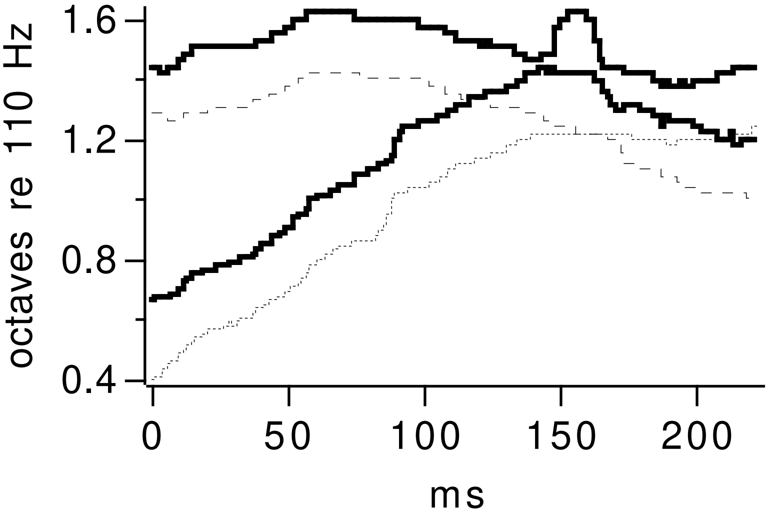

Fig. 2 F_0 estimates as a function of time, showing the breakdown of the algorithm when the f_0 tracks cross. Female speaker, tokens b3 and b7.

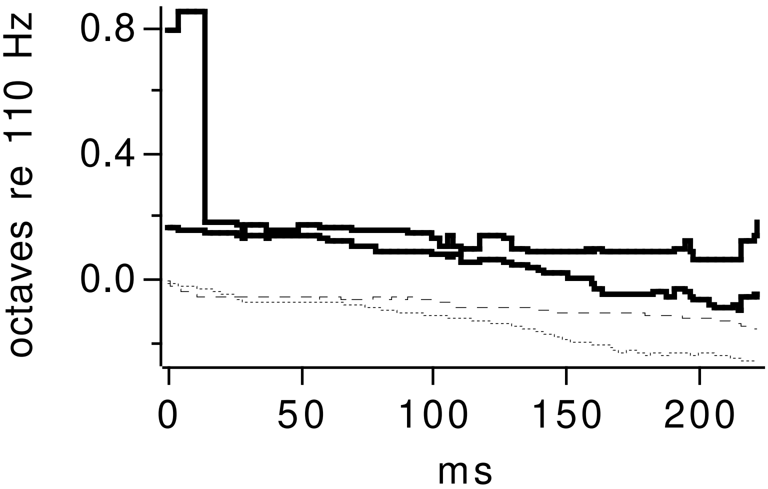

Fig. 3 F_0 estimates as a function of time, illustrating the behavior of the algorithm when f_0 tracks are close. Male speaker, tokens a0 (as in fig. 1) and a3.

Figure 4 shows a histogram of the differences between estimates and their closest target values, pooled over the whole data set. Note the log scale: on a linear scale the histogram is too sharp for interpretation. 90% of the estimates made by the algorithm are within 3% of a target value, and 65% are within 1%.

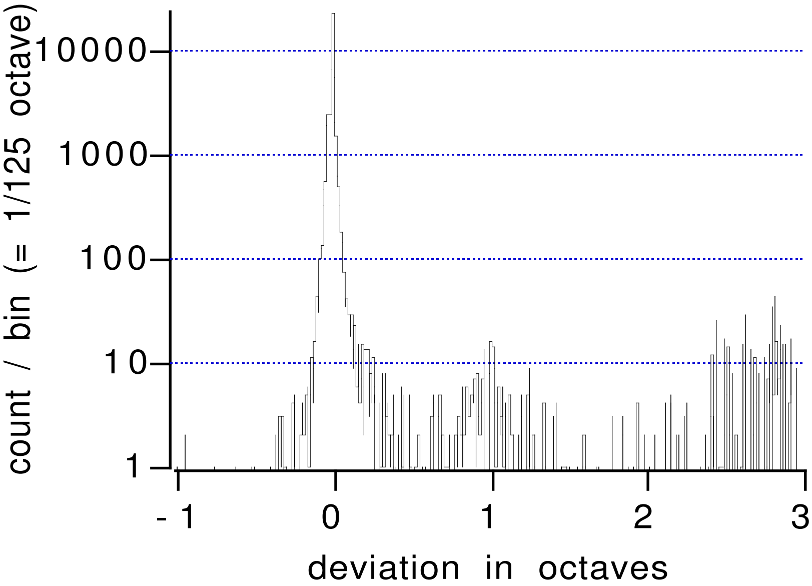

Fig. 4 Histogram of the deviation in octaves of f_0 estimates made by the mixed speech algorithm from the closest target value. Note the log scale.

The following table shows another measure of reliablity, the mean error magnitude (in % octave), for the different data subsets at two sampling rates: the original 20 kHz rate and a 40 kHz rate obtained by linear interpolation:

data set:	mean error (20 kHz)	mean error (40 kHz)
male/male	6.3	6.1
female/female	7.0	6.3
male/female	8.2	7.3

The mean error for female/female at 40 kHz is the same as that for male/male at 20 kHz. This suggests that the slight disadvantage of female/female over male/male can be ascribed to the greater effect at higher frequencies of the limited sampling resolution. The slight disadvantage of the male/female condition relative to the others may be due to the fact that there is less overlap in the search ranges for both voices, and therefore a wider overall search range. Apart from these, there are no major differences between the conditions. In particular, the algorithm does not require the f_0 tracks to be in different registers, contrary to other algorithms (see for example Weintraub 1985).

CONCLUSION

In summary, the mixed voice fundamental frequency estimation algorithm appears to be quite successful in finding the fundamental frequencies of both voices. The restrictive conditions of the evaluation must be stressed: only "clean" voiced speech was used (according to a criterion that eliminates 25% of voiced speech), and the search range was restricted to 3 octaves. On the other hand, the algorithm performs the task on a frame-by-frame basis, using only local information: continuity, or voice register constraints could further enhance reliability.

In practical applications such as voice separation, a mixed voice f_0 estimation algorithm must recognize and cope with situations where only one speech track is present or voiced. It must also be able to follow f_0 tracks when they cross (the present algorithm makes no such attempt). These are subjects for future research.

The algorithm is time consuming, but easily parallelizable. Search techniques smarter than exhaustive search can also be used.

ACKNOWLEDGEMENTS

Part of this work was carried out when the author was staying at the ATR Auditory and Visual Perception Research Laboratories (Japan). The author wishes to thank the Centre National de la Recherche Scientifique (CNRS) for leave of absence, and the ATR Auditory and Visual Perception Research Laboratories for their kind hospitality and support. Thanks are also due to the Institut de Recherche et de Coordination Acoustique-Musique (IRCAM) for access to their facilities. Eric Bateson, William Campbell, Tatsuya Hirahara, Dominic Massaro, Steve McAdams, Ray Meddis, Yoh'ichi Tohkura and Minoru Tsuzaki made useful comments on an earlier report (de Cheveigné 1990).

BIBLIOGRAPHY

Assmann, P. F. and Q. Summerfield. (1990). "Modeling the perception of concurrent vowels: vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88, pp. 680-697.

Cherry, E. C. (1953). "Some experiments on the recognition of speech with one and with two ears," J. Acoust. Soc. Am. 25, pp. 975-979.

de Cheveigné, A. (1990). "F0 estimation from mixed speech," ATR Auditory and Visual Perception Research Laboratories technical report.

Duda, R. O. and R. F. Lyon. (1990). "Correlograms and the separation of sounds," Asilomar annual conference on signals, systems and computers.

Frazier, R. H., S. Samsam, L. D. Braida and A. V. Oppenheim. (1976). "Enhancement of speech by adaptive filtering," IEEE ICASSP, pp. 251-253.

Hess, W. (1983). Pitch determination of speech signals (Springer-Verlag, Berlin). 698 p.

Kitamori, S., T. Harada and H. Kawarada. (1984). "Auditory nerve impulse cross-correlation model for separation of mixed speech," Proceedings of the ASJ commitee on Hearing Research. H-84-6, pp. 1-6 (in Japanese).

Kuwabara, H., Y. Sagisaka, K. Takeda and M. Abe. (1989). "Construction of ATR Japanese speech database as a research tool," ATR Interpreting Telephony Research Laboratories technical report (in Japanese).

Meddis, R. and M. J. Hewitt. (1990). "Modelling the identification of concurrent vowels with different fundamental frequencies," Submitted for publication.

Nagabuchi, H., T. Kobayashi and H. Yamamoto. (1979). "Speech enhancement and suppression in mixed speech," Transactions of the IECE (Japan) 62, pp. 627-634 (in Japanese).

Palmer, A. R. (1988). "The representation of concurrent vowels in the temporal discharge patterns of auditory nerve fibers," in Basic issues in hearing, edited by H. Duifhuis, J. W. Horst and H. P. Wit (Academic Press, London), pp. 244-251.

Palmer, A. R. (1990). "The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochler-nerve fibers," J. Acoust. Soc. Am. 88, pp. 1412-1426.

Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. 60, pp. 911-918.

Ross, M. J., H. L. Shaffer, A. Cohen, R. Freudberg and H. J. Manley. (1974). "Average magnitude difference function pitch extractor," IEEE Trans. ASSP 22, pp. 353-362.

Scheffers, M. T. M. (1983). "Sifting vowels". Thesis, Gröningen University.

Stubbs, R. J. and Q. Summerfield. (1988). "Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 84, pp. 1236-1249.

Stubbs, R. J. and Q. Summerfield. (1990). "Algorithms for separating the speech of interfering talkers: evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 87, pp. 359-372.

Weintraub, M. (1985). "A theory and computational model of auditory monaural sound separation". Thesis, Stanford University, 158p.

Weintraub, M. (1986). "A computational model for separating two simultaneous sounds", IEEE ICASSP, Tokyo, 1, pp. 3.1.1-3.1.4.