SPEECH F0 EXTRACTION BASED ON LICKLIDER'S PITCH PERCEPTION MODEL

Alain de Cheveigné

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

ABSTRACT

According to a pitch perception model proposed by Licklider [1, 2, 3], time-domain patterns of activity in nerve channels coming from the cochlea undergo autocorrelation analysis in the auditory nervous system. We examine whether this model can be adapted to the task of speech f0 estimation, and in particular what benefit the filter-bank processing stage can bring to a fundamental period estimation algorithm. Results show an improvement in reliability over the same algorithm applied directly to the speech signal.

1. INTRODUCTION

1.1. Perception models applied to f0 extraction

A large number of speech f0 estimation algorithms have been proposed [4]. Some are purely signal processing methods, others derive from models of speech production or perception. While they mostly give similar results on clearly periodic voiced speech, some may fail or give doubtful results on less periodic portions [5]. Aperiodicity of voiced speech can be due in some cases to severe irregularity in occurrence of glottal pulses. In such cases it is impractical to define f0 in terms of *production* (as the inverse of the interval between glottal pulses), and it may seem preferable to define it instead in terms of *perception* (pitch).

Several perception-based methods have been proposed [6, 7, 8], most of which are based on the pitch perception theories of Goldstein or Terhardt [9, 10, 11]. The general principle shared by these models is that pitch is determined from a spectral pattern by searching for a common subharmonic of major spectral components. The spectral pattern is presumably produced by peripheral analysis in the cochlea, and the matching of subharmonics carried out at a more central stage. Spectral pattern matching theories are being questioned of late, because physiological data support alternative theories that assume that pitch derives instead from the *periodicity of neural discharges*.

1.2. Licklider's model of pitch perception

Licklider [1, 2, 3] proposed a model according to which each channel within the auditory nerve is processed by an autocorrelation mechanism. The result of this processing is a pattern of neural activity over the dimensions of *frequency* (inherited from cochlear filtering) and *lag* (implemented as nerve conduction or synaptic delay). In response to a periodic stimulus such as voiced speech, a ridge appears spanning frequency at a lag equal to the period. The position of this ridge is the cue to pitch. Licklider's ideas have been developped recently by other authors [12, 13, 14, 15, 16]. Autocorrelation, as used in Licklider's model, does not *require* a filtering stage: it can be performed directly on the raw speech signal [4]. This raises a question: what might be the advantage of peripheral filtering for pitch perception? One can imagine several possible answers:

a) The signal-to-noise ratio or the periodicity might be better within a restricted group of channels. [17][18].

b) Small differences of phase from period to period can result in large differences in wave shape, causing a comparison method such as autocorrelation to fail. Filtering might reduce such interaction.

1.3. Applying the model to f0 extraction

The aim of this paper is to verify experimentally whether *splitting a speech signal over a filter bank* offers any advantage for speech f0 extraction. It is important to stress that we do not aim to reproduce all aspects of the perception model in the extraction method. The perceptual quality called pitch is not the same object as speech fundamental frequency (often also called pitch) and the tasks of extracting the former or perceiving the latter are not equivalent.

2. METHODS

2.1. Database

Data was taken from an f0 database developped at ATR [19, 20]. The speech was sampled at 12 kHz with 16 bit resolution, and labeled for pitch by a crude cepstrum method followed by manual correction. The database contains 500 sentences, read by one male speaker, of which 20 "difficult" sentences were selected and carefully re-labeled by hand. The sentences comprise approximately 19000 voiced frames at a 400 Hz frame rate. The f0 values cover a 2-octave range centered on about 125 Hz.

2.2. AMDF

All experiments are base on the Average Magnitude Difference Function (AMDF) method [21]. The AMDF is defined as:

$$\mathrm{AMDF}(lag) = \int_{\mathrm{window}} |S(t) - S(t + lag)|\,dt$$

The lag at the first major dip indicates the period. The AMDF produces as a by-product a parameter that can be interpreted as a *measure of periodicity*. This is defined as:

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

The periodicity can be used as a measure of "confidence" in the period value produced by the AMDF algorithm, and also to select channels of high periodicity.

2.3. Evaluation

The AMDF search was constrained to search within 30% of the period specified in the database. The lag at this minimum, the periodicity measure, and an error code are output for each frame. The error code indicates whether the algorithm would have been successful without constraint. It distinguishes subharmonic errors which are *not counted as errors* in this paper. A "baseline" record of these parameters was derived for the database using standard AMDF. Evaluation was done by frame-to-frame comparison to this baseline. Care was taken to preserve the alignment of processed data: signal smoothing was performed with symmetrical windows, and the outputs of the revcor filters (see below) were shifted in time and phase-adjusted so that the peaks of the envelope and fine time structure of their impulse response coincided with the time origin.

2.4. Revcor filter bank

The experiments use a filter bank program [22] that approximates peripheral auditory filters as "revcor" (or "gammatone") filters, defined by their impulse response:

$$h(t) = A(t - T_l)^{\upsilon} \exp(-(t - T_l) / T_f) \sin(2\pi F(t - T_l))$$

where F is the characteristic frequency, T_l is a latency, T_f is a time constant of decay, and ν is a factor that governs the "symmetry" of the impulse response. The bandwidth parameter was derived from psychoacoustical masking data [23]. Physiological data indicate bandwidths up to three times larger [24, 25]; this factor is explored in the experiments. Bandwidths were set at 1 (standard), 2, 4 and 8 ERB (Equivalent Rectangular Bandwidths) [23]. The filter produces 25 channels uniformly spaced at 1 ERB intervals from 40 Hz to 4000 Hz.

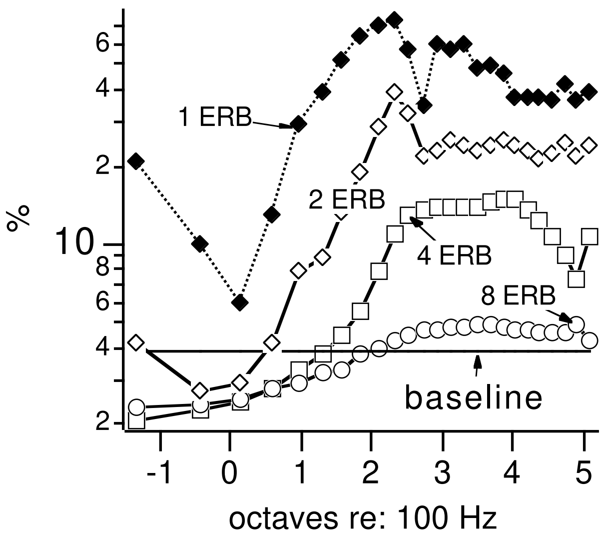

Fig. 1. Error rate as a function of center frequency for various channel bandwidths measured in ERBs.

3. EXPERIMENTS

3.1. Baseline

The error rate of "vanilla" AMDF over the database is 3.84%.

3.2. Individual revcor channels.

The error rates are displayed in Fig. 1 for several bandwidth settings. The rates at 1 ERB bandwidth are very high (around 50%), for other bandwidths they are more reasonable. Rates are lower than baseline in low-frequency channels, and higher in high frequency channels. The rates at 8 ERB are not very different from baseline, a result which was to be expected given the rather wide filters.

3.3. Half-wave rectification and low-pass filtering.

A possible cause for less good rates in high frequency channels is that it is harder to "register" the fine waveform structure of successive periods. In the auditory system much of this detail is lost, because of the fall-off of synchrony from 1 to 5 kHz [26], an effect similar to smoothing. To check the possible benefit of this effect, the revcor channel outputs were half-wave rectified and smoothed by convolution with a 20 ms rectangular window (first zero at 500 Hz). Results show an improvement in high-frequency channels, and a slight degradation in low-frequency channels, perhaps because of the loss of information that accompanies half-wave rectification.

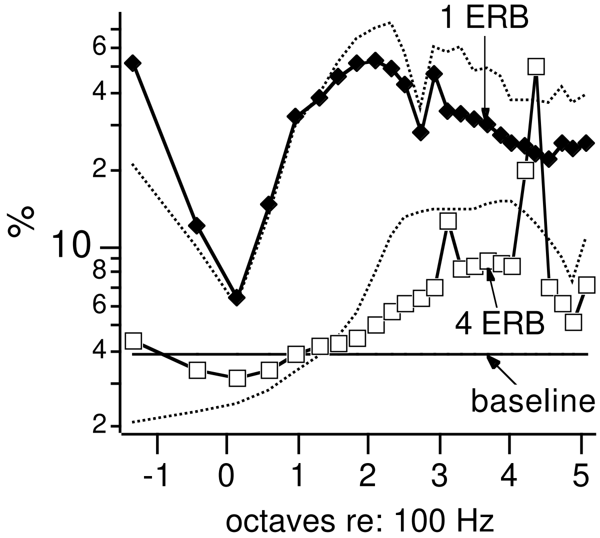

Fig. 2. Error rates for half-wave rectified revcor filter outputs. Dotted lines: rates for raw outputs.

3.4. Cross-channel integration

There are many ways of combining patterns. Here we report a few:

- *addition of AMDFs*

The AMDF patterns for all channels are added before searching for the minimum that indicates the period. Error rate, for 1 ERB bandwidth, is 2.9 %

- *addition of AMDFs of amplitude normalized channels*

The revcor filter channels are amplitude normalized (by division by the mean magnitude over a centered window) to give each channel the same weight. Error rate for 1 ERB bandwidth is 5.15 %.

- *addition of AMDFs of half-wave rectified, smoothed channels*

Error rate for 1 ERB bandwidth is 2.7 %.

4. DISCUSSION

At a bandwidth of 1 ERB the error rates are high, probably because resolution of partials prevents interaction at the fundamental. Rates are much lower at wider bandwidths, particularly for low frequency channels, which suggests that periodicity information is somehow "better" in these channels. This interpretation is confirmed by results for low-pass filtered speech (table 1).

Table 1: error rates for various degrees of smoothing:

window size:	10 ms	20 ms	40 ms	80 ms
zero at:	1 kHz	500 Hz	250 Hz	125 Hz
error rate (%):	3.19	2.44	2.74	3.96

Given this simple result, one might be tempted to apply low-pass filtering systematically. This would be unwise for a number of reasons. For one, the optimum cutoff frequency depends on the pitch range, and a good setting in one case might be disastrous in others. For another, some applications call for pitch extraction of high-pass filtered speech (such as telephone speech), in which case there is evidently no benefit in low-pass filtering. A more robust strategy appears to be to combine information across channels. Simple addition of AMDF patterns yield 2.9 % errors for a 1 ERB bandwidth. This is in striking contrast with the rates obtained in individual channels (Fig. 1). Better still is the rate for summed AMDF patterns of half-wave rectified, smoothed channels (2.7% for 1 ERB bandwidth). Uniform weights for all channels, as obtained by amplitude normalization, proved disappointing (5.15% for 1 ERB bandwidth).

CONCLUSION

An f0 extraction method based that splits the speech signal over a filter-bank before calculating the AMDF within each channel and combining the patterns improves reliablity of the AMDF method. Future work will examine more sophisticated schemes, such as weighting each channel according to its periodicity measure. More complex algorithms can also be used, such as the channel selection algorithms used by some multiple-source separation models [27, 28].

ACKNOWLEDGMENTS

Part of this work was carried out at ATR Interpreting Telephony Research Laboratories, under a fellowship awarded by the European Communities STP programme. The author wishes to thank ATR for its hospitality, and the CNRS for leave of absence. Special thanks is due to John Holdsworth and Roy Patterson who made available the revcor filter software, and to IRCAM for use of their facilities.

BIBLIOGRAPHY

[1.] Licklider, J. C. R. (1956), "Auditory frequency analysis", *Information theory*, Cherry ed. Butterworth: London, 253-268.

[2.] Licklider, J. C. R. (1959), "Three auditory theories", *Psychology, a study of a science*, Koch ed. McGraw-Hill: 41-144.

[3.] Licklider, J. C. R. (1962), "Periodicity pitch and related auditory process models", International Audiology. *1*, 11-36.

[4.] Hess, W. (1983), *Pitch determination of speech signals*, Springer-Verlag: Berlin. Pages.

[5.] Hedelin, P. and D. Huber (1990), "Pitch period determination of aperiodic speech signals", *IEEE-ICASSP*, 361-364.

[6.] Duifhuis, H., L. F. Willems and R. J. Sluyter (1982), "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception", JASA. 1568-1580.

[7.] Hermes, D. J. (1988), "Measurement of pitch by subharmonic summation", JASA. *83*, 257-264.

[8.] Scheffers, M. T. M. (1983), "Sifting vowels",

[9.] Goldstein, J. L. (1973), "An optimum processor theory for the central formation of the pitch of complex tones", JASA. *54*, 1496-1516.

[10.] Terhardt, E. (1974), "Pitch, consonance and harmony", JASA. *55*, 1061-1069.

[11.] de Boer, E. (1977), "Pitch theories unified", *Psychophysics and physiology of hearing*, Evans and Wilson ed. Academic: London, 323-334.

[12.] Moore, B. C. J. (1982), *An introduction to the psychology of hearing*, Academic Press: London. Pages.

[13.] van Noorden, L. (1982), "Two channel pitch perception", *Music, mind, and brain*, Clynes ed. Plenum press: London, 251-269.

[14.] Lyon, R. (1984), "Computational models of neural auditory processing", *IEEE ICASSP*, 36.1.(1-4).

[15.] de Cheveigné, A. (1986), "A pitch perception model", *Proc. IEEE ICASSP*, 897-900.

[16.] Meddis, R. and M. Hewitt (1988), "A computational model of low pitch judgement", *Basic issues in hearing*, Duifuis, Horst and Witt ed. Academic: London, 148-153.

[17.] Fujimura, O. (1968), "An approximation to voice aperiodicity", IEEE Trans. Audio and Electroacoustics. *16*, 68-72.

[18.] Rodet, X., P. Depalle and G. Poirot (1988), "Diphone sound synthesis based on spectral envelopes and harmonic/noise excitation functions.", Proceedings of the ICMC, Köln (RFA), 313-321.

[19.] Kuwabara, H., Y. Sagisaka, K. Takeda and M. Abe (1989), "Construction of ATR Japanese speech database as a research tool", ATR technical report TR-I-0086.

[20.] Abe, M. and H. Kuwabara (1989), "Pitch frequency database on continuous speech", ATR technical report TR-I-0078.

[21.] 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*, 353-362.

[22.] Holdsworth, J., I. Nimmo-Smith, R. D. Patterson and P. Rice (1988), "Implementing the GammaTone filter bank",

[23.] Moore, B. C. J. and B. R. Glasberg (1983), "Suggested formulae for calculating auditory filter bandwidths and excitation patterns", JASA. *74*, 750-753.

[24.] Carney, H. and T. C. T. Yin (1988), "Temporal coding of resonances by low-frequency auditory nerve fibers: single fiber responses and a population model.", J. Neurophysiol. *60*, 1653-1677.

[25.] de Cheveigné, A. (1990), "Experiments in pitch extraction.", ATR Technical report TR-I-0103, 39p.

[26.] Johnson, D. H. (1980), "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones", JASA. *68*, 1115-1122.

[27.] Lyon, R. F. (1983-1989), "A computational model of binaural localization and separation", *Natural computation*, Richards ed. MIT Press: Cambridge, Mass, 319-327.

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