SEPARATION OF SPEECH FROM SIMULTANEOUS TALKERS

R.I. Damper, J.R. Thorpe and C.H. Shadle
Department of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK

ABSTRACT

The separation of speech from two simultaneous talkers is a problem of some practical and theoretical importance. We describe a prototype separation system based on harmonic selection using comb filters. Hermes' subharmonic spectrum method is used to produce a number of (weighted) pitch estimates, with pitch tracks for the two talkers then found by constrained dynamic programming. The system has successfully separated composite male/female /hVd/ tokens but performance is currently rather variable.

INTRODUCTION

The separation of a target speech signal from contaminating, competing signals is a problem of some significance, having applications to improved speech recognition and signal-processing hearing aids. An especially interesting instance of the problem arises when the (single) contaminating source is the speech of another talker. Not only is this a common situation in practice, but separation is likely to be maximally difficult since the target and contaminating signals will share obvious similarities.

Early approaches to this problem [1] were monaural, estimating the fundamental frequency (or 'pitch') of each talker (f_0^1 and f_0^2 respectively), then selecting components of the frequency spectrum and assigning them to a talker according to their harmonic relation to the estimated pitch(es). This *harmonic selection* method assumes that the speech of at least one of the two talkers is voiced, and requires f_0^1 and f_0^2 to be well spaced so that it is obvious which talker is which.

Harmonic selection can be viewed as implementing one of the perceptual grouping principles advanced by Bregman [2], whereby human listeners are able to aggregate auditory features arising from distinct sound sources to effect separation. Other putative grouping principles are based on onset and/or offset synchrony of features, a common rate of amplitude modulation, and cues suggesting a common spatial origin.

Clearly, any implementation of harmonic selection is critically dependent upon a robust pitch detection algorithm (PDA) but most PDAs assume a single voice only [3,4]. More recent work on talker separation [5,6,7] has, therefore, focussed on improved PDAs. However, given that a common spatial origin is likely to be important to grouping, and thereby separation, attention has also been paid to binaural techniques [7,8]. Denbigh and Zhao [7] state that the major advantage of their binaural technique is the ability to recover from talker-allocation errors when f_0^1 and f_0^2 tracks cross.

We describe here the implementation of a prototype monaural separation system which has been successfully applied to the two-talker problem. In the next section, we detail the speech data employed in this study. We then describe the use of Hermes' subharmonic spectrum (SHS) pitch detection algorithm [9] to obtain several weighted estimates of f_0,

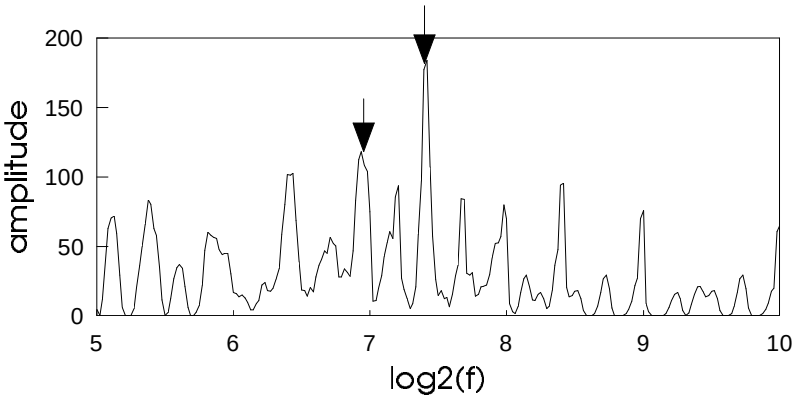

Figure 1. Subharmonic spectrum of typical frame of composite token shown here for $\log_2 f = 5$ to 10, i.e. 32 to 1024 Hz. Actual fundamental frequencies for the two talkers are shown arrowed.

without consideration of talker identity at this stage. A dynamic-programming (DP) tracking algorithm is then described. This is used to correct pitch errors and to allocate optimal f_0 tracks to each of the two talkers. Results of separation using comb filters are then detailed.

SPEECH DATA

The speech data used in this study were a subset of those recorded by Deterding [10], consisting of /hVd/ tokens spoken by 3 male and 3 female adults and sampled at 10 kHz. A small number of composite tokens was then formed by adding (arbitrarily selected) pairs of male and female tokens. Male/female pairs were chosen to minimise problems of crossing pitch tracks – since the present, prototype implementation is monaural.

Processing was based on frames of 512 samples with 50% overlap. Each frame was multiplied by a Hanning window, padded with a further 512 zeros, and a 1024-point FFT taken. The resulting frequency resolution is, therefore, 9.77 Hz.

SHS ALGORITHM

Hermes' SHS algorithm [9] is an improved version of the harmonic compression PDAs of Schroeder [11] and Noll [12]. These rely on compressing the frequency scale of a spectral representation by integer (harmonic) factors and then taking either the product or the sum of the compressed representations, e.g. Noll's harmonic sum spectrum is defined as:

$$S(f) = \sum_{k=1}^{K} |F(kf)|^2$$

where $F(f)$ is the Fourier spectrum and there are $(K-1)$ compressions. The fundamental f_0 then appears as a peak in the product or sum spectrum, as there is consistent reinforcement of the fundamental by the compressed harmonics.

The problem with these algorithms is that there is a loss of data when used with sampled signals, since certain of the sample points in the compressed spectra fall between those in the original ($k = 1$) spectrum. This severely limits the value of K which can be employed (to about 5). The SHS algorithm avoids this problem by substituting harmonic compression on a linear frequency scale by harmonic shift on a logarithmic scale. Also, the amplitude spectrum (rather than the power spectrum) is used, with decreasing weight given to the more compressed spectra:

$$S(\log_2 f) = \sum_{k=1}^{K} w(k)|F(\log_2 f + \log_2 k)|$$

where here $w(k) = 0.84^{k-1}$ and $K = 9$.

Since the linear-to-log frequency conversion results in logarithmically-spaced sample points, the spectrum is resampled

by cubic spline interpolation at 48 points per octave after conversion. There is also a broadening of spectral peaks at lower frequencies; accordingly, peaks are thinned to a constant width of 3 samples in the log frequency domain. Figure 1 shows a typical subharmonic spectrum with the actual f_0^1 and f_0^2 marked by arrows.

Since even the best PDA will make frame errors, we do not attempt to identify f_0^1 and f_0^2 uniquely at this stage. Rather, the SHS algorithm produces six weighted estimates of possible fundamental for later DP pitch tracking as follows. The 3 largest peaks of the SHS are selected and weighted 1, 2 and 3 respectively. The largest peak (weighted 1) is then assumed to correspond to f_0 for the dominant talker. This estimate of f_0 and its harmonics are then used to subtract corresponding peaks from the thinned Fourier spectrum, and the SHS algorithm re-run to produce 3 new f_0 estimates, again weighted 1,2,3. As a consequence of the use of a log frequency scale, the resolution of the f_0 estimates is nonlinear (being $48\log_2 f$).

No distinction is made between voiced and unvoiced speech, both being treated identically.

DP PITCH TRACKING

By maintaining multiple candidate f_0 values, improved pitch estimates can be obtained by dynamic-programming (DP) tracking. We use the method described by Ney [13] which performs a DP optimisation constrained by a (weighted) 'measurement' cost and a 'smoothness' cost.

The input to the DP algorithm is an $n \times m$ time-frequency matrix, where n is the number of possible f_0 values and m is the number of frames in the composite token. Because f_0 is assumed to lie between 32 and 512 Hz, values ≤ 32 are considered to be 0 while values ≥ 512 are considered to be 512 Hz. Hence, there are

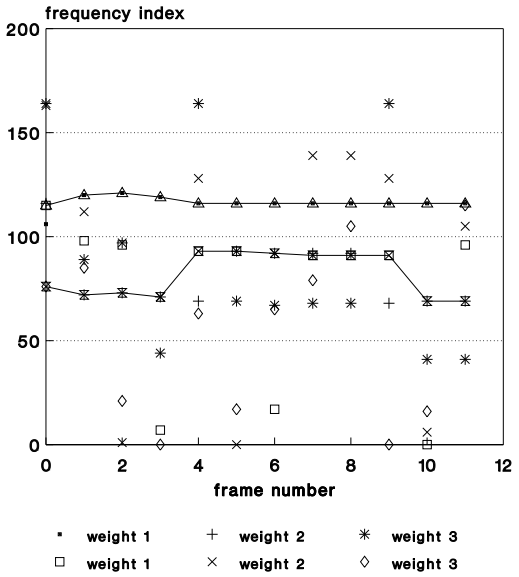

Figure 2. Time-frequency matrix for typical composite token and the pitch tracks (solid lines) for the two talkers found by dynamic programming.

$n = 48(\log_2 512 - \log_2 32) = 192$ 'frequency indices', according to the logarithmic resolution of the PDA. The cells of the matrix contain the measurement cost, and are initialised to a high value, W_{ini}. The weights of the 6 f_0 estimates ($W \in \{1, 2, 3\}$) from the SHS algorithm are then entered in the appropriate cells.

The smoothness cost, D, was implemented as the absolute difference between frequency indices for consecutive frames, so penalising departure from a constant pitch value. The total cost is then the linear combination $W + \alpha D$.

With W_{ini} and α set empirically (at 100 and 0.2 respectively), the DP algorithm was applied to the matrix to find the optimal pitch track for one of the talkers. The values in cells on this track were then replaced by W_{ini} and the algorithm re-run to find the optimal pitch track for the second speaker. This is shown in Figure 2 for a typical composite token.

It is difficult to validate the pitch tracks found. However, use of a commercially-available speech analyser (Kay CSL) gave excellent agreement for one speaker and reasonable agreement for the other.

SEPARATION ALGORITHM

First, the Fourier spectrum is differentiated to find all its maxima, which are listed. The separation algorithm then takes the larger of f_0^1 and f_0^2, and uses this to calculate tentative values for the harmonic frequencies. These are then matched to the list of maxima; if there is no peak at the exact harmonic value, the points either side are checked to see if they are maxima.

Each harmonic peak thus found becomes the centre of one tooth of a comb filter. Each tooth is 5 FFT points wide, and has a Hanning window shape. Multiplication of the Fourier spectrum by the comb filter response then yields a frame of separated data corresponding to the higher f_0. Peaks allocated to this speaker are then deleted from the list of maxima, and the process repeated for the lower f_0.

When this has been done for all frames, separated tokens are produced by overlap-add re-synthesis.

CONCLUSIONS

As judged by informal listening, the prototype separation system works extremely well for some of the composite tokens but less well for others. Separation is better for female than for male talkers – the male separated tokens being more affected by cross-talk. Given the relatively small database used, this may simply reflect lower pitch variation among the female talkers which results in more accurate pitch tracking.

REFERENCES

[1] Parsons, T.W. (1976) "Separation of speech from interfering speech by means of harmonic selection", *Journal of the Acoustical Society of America*, vol. 60, pp. 911–918.

[2] Bregman, A.S. (1990) *Auditory scene analysis*, Cambridge, MA: MIT Press.

[3] Hess, W. (1983) *Pitch determination of speech signals: algorithms and devices*, Berlin: Springer-Verlag.

[4] Hermes, D.J. (1993) "Pitch analysis", in *Visual representations of speech signals*, M. Cooke, S. Beet and M. Crawford (eds.), Chichester, UK: Wiley, pp. 3–25.

[5] Weintraub, M. (1987) "Sound separation and auditory perceptual organization", in *The psychophysics of speech perception*, M.E.H. Schouten (ed.), Dordrecht: Martinus Nijhoff, pp. 125–134.

[6] Stubbs, R.J. and Summerfield, Q. (1990) "Algorithms for separating the speech of interfering talkers: Evaluation with voiced sentences, and normal-hearing and hearing-impaired listeners", *Journal of the Acoustical Society of America*, vol. 84, pp. 1236–1249.

[7] Denbigh, P.N. and Zhao, J. (1992) "Pitch extraction and separation of overlapping speech", *Speech Communication*, vol. 11, pp. 119–125.

[8] Banks, D. (1993) "Localisation and separation of simultaneous voices with two microphones", *IEE Proceedings Part I*, vol. 140, pp. 229–234.

[9] Hermes, D.J. (1988) "Measurement of pitch by subharmonic summation", *Journal of the Acoustical Society of America*, vol. 84, pp. 257–264.

[10] Deterding, D.H. (1990) *Speaker normalisation for automatic speech recognition*, DPhil Thesis, University of Cambridge, UK.

[11] Schroeder, M.R. (1968) "Period histogram and product spectrum: new methods for fundamental-frequency measurement", *Journal of the Acoustical Society of America*, vol. 43, pp. 829–834.

[12] Noll, A.M. (1970) "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate", *Proceedings of Symposium on Computer Processing in Communications*, New York: University of Brooklyn Press, pp. 779–798.

[13] Ney, H. (1983) "Dynamic programming algorithm for optimal estimation of speech parameter contours", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 13, pp. 208–214.