CEPSTRAL AND SPECTRAL APPROACHES TO SEPARATING CONCURRENT VOICES

Richard J Stubbs and Quentin Summerfield

M.R.C Institute of Hearing Research, University Park, Nottingham

INTRODUCTION

Is it possible to develop a computer algorithm to separate the voices of competing speakers? Normal-hearing listeners segregate speech sources routinely by using binaural and monaural cues, visual information and acoustic, linguistic and articulatory constraints (1). This paper addresses one aspect of this formidable array of abilities, that of monaural segregation of periodically-excited voice sources.

Why might a computer algorithm for segregating voices be of interest? The sound of competing speakers is one of the most common forms of noise. It produces a disproportionate amount of interference in relation to its energy because the noise rejection task involves segregating signals with similar spectro-temporal properties. It may be possible to improve the performance of a normal-hearing listener by pre-processing the signal. Moreover, such pre-processing could be of significantly greater benefit to listeners with a hearing impairment of cochlear origin, for whom suprathreshold limitations on spectro-temporal resolution (e.g. (2)) make the task of separating speech from noise particularly difficult (e.g. (3)). A voice segregation algorithm would be an attractive inclusion in a signal-processing hearing-aid. Production of an aid of this processing power lies several decades away. However, assessment of the physical principles, feasibility and cost benefit of the elements of such a system are within the scope of present-day technology.

A computational model of auditory monaural sound separation has been described by Weintraub (1). We have evaluated two signal-processing algorithms of limited applicablity, which nonetheless might form part of a segregation strategy. They are designed to separate the voices of competing talkers speaking with similar intensities. The algorithms operate only on voiced speech, that is those portions of speech which have a regular harmonic structure in the frequency spectrum. For successful separation there must be a difference in fundamental frequency between the voices. Although they constitute only about 50% of natural speech, the voiced portions of speech are the obvious place to start voice separation. They form the most intense parts of the signal and are robustly coded against interference because of the high inter-correlation between frequency components created by the harmonic structure.

The two algorithms illustrate contrasting approaches to voice separation. One is a noise-reduction process. It attenuates the harmonic excitation of the unwanted speaker in a cepstrum-like representation of the signal. The other is a signal-extraction processs. It uses the method of harmonic selection (4) to group together harmonic peaks in the frequency spectrum that appear to constitute a voice and uses that information to reconstruct the voice. The second technique is computationally more expensive but has greater potential and generality, since it exploits knowledge of the characteristics of speech signals rather than of a specific noise.

The algorithms were evaluated both subjectively and in formal listening tests. The tests addressed three questions:

1. Does a normal-hearing listener derive any benefit from pre-processing by the algorithms?
2. Does a hearing-impaired listener derive any benefit?
3. What is the relative performance of the two methods?

PROCESSING ARCHITECTURE

All processing was done on a DEC PDP11/60 minicomputer in greater than real-time. Both processing methods have the same overall structure. A signal containing two voices is low-pass filtered at 4.25kHz and digitised at a 10kHz sampling rate with 12-bit amplitude quantisation. The digitised signal is divided into 1/2-overlapping, 51.2ms, hanning-windowed segments which are processed sequentially. The segment duration is large enough to create an amplitude spectrum in which the harmonic structure of the signal is resolved, whilst being short compared to many of the dynamic properties of speech. The amplitude and phase spectrum of each segment are computed from an FFT. The amplitude spectrum, containing both signals, is processed to separate it into the two constituent spectra. The two separated amplitude spectra are each combined with the original phase spectrum and are transformed back to the time-domain by a reversal of the original segmentation procedure.

CEPSTRAL FILTERING

The first method of processing the amplitude spectrum employs homomorphic filtering (5). The rationale behind such a scheme, can be illustrated by considering a single speaker. If speech can be considered to be stationary over the duration of a 51.2ms segment, then the amplitude spectrum S(f) is the product of the excitation e(f), in this case a harmonic series, and the vocal tract transfer function v(f) (fig(1a)).

$$S(f)=e(f).v(f) \; ..... \quad (1)$$

In the logarithmic spectrum the product becomes additive, creating a linear superposition of excitation and envelope (fig(1b)). If log S(f) is Fourier transformed to produce a spectrum of the logarithmic spectrum fig(1c), then the speech envelope is modelled predominantly by slowly varying oscillations and the harmonic series by a rapidly varying component. These appear as a broad peak near the origin and a sharper 'pitch' peak in fig(1c). The new representation is very similar to a cepstrum* and will be referred to as such. If two harmonic voice sources are present in the original signal, the simple superposition in the logarithmic spectrum no longer applies. However, the cepstrum has the same general appearance, and if the sources have different

---

* A true cepstrum is defined as the INVERSE Fourier transform of the log spectrum. In the filtering operation on the spectrum, we use the sequence FFT, FILTER, IFFT rather than the true cepstral sequence IFFT, FILTER, FFT.

fundamental frequencies, two pitch peaks are observed, corresponding to the two constituent harmonic series. Removing the pitch peak of one voice can enhance the voice whose pitch peak remains. In the processing an FFT is performed on all of the logarithmic amplitude spectrum excepting the D.C. term and the magnitude and phase of the transform are computed. Attenuation is applied to the cepstrum magnitude. In its present implementation the algorithm must be told which region of the cepstrum is to be attenuated. It is assumed that the competing voices occupy non-overlapping pitch ranges.

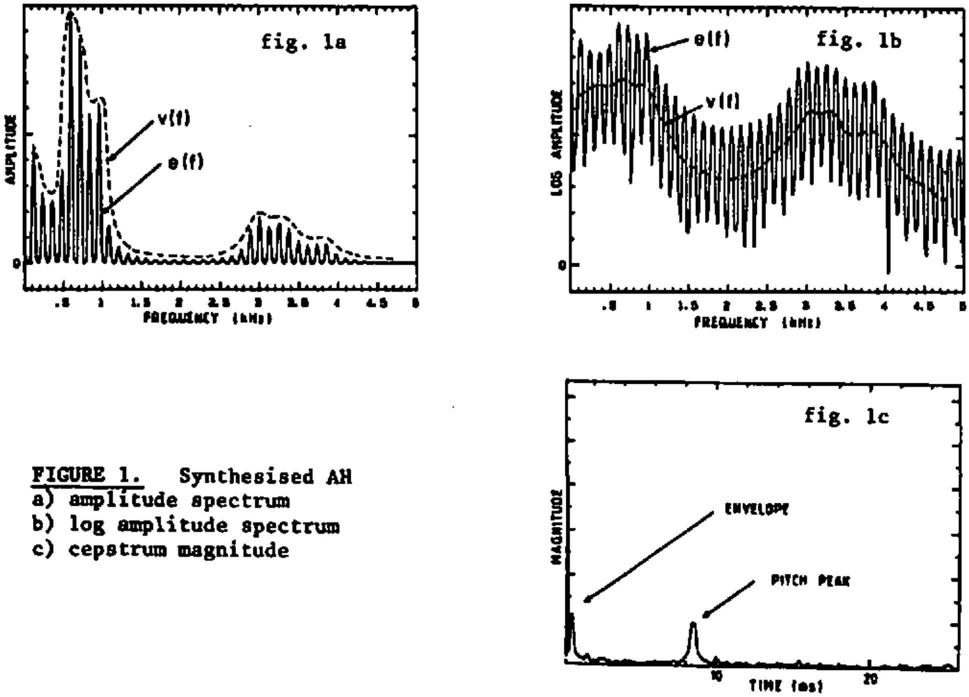

FIGURE 1. Synthesised AH
a) amplitude spectrum
b) log amplitude spectrum
c) cepstrum magnitude

Removing the cepstral pitch peak of a voice attenuates the harmonic excitation of that voice. For highly harmonic sources, such as the synthesised vowels used in the perceptual experiments reported below, pitch peaks also occur in the cepstrum at submultiples of the fundamental frequency f0, f0/2..... . Where the amplitude of the f0/2 peak was significant it was also attenuated. In examples using natural speech, where the overall region containing the pitch peak of the unwanted speaker was known, that whole region was zeroed.

An example of cepstral filtering separating two concurrent, synthesised vowels is illustrated in the frequency domain in fig(2). The two vowels, AH and EE, have fundamental frequencies of 120Hz and 151Hz respectively figs(2a,2b). In the combined spectrum (fig(2c)) the two vowels are heavily overlapped. In the post filtering spectra (figs(2d,2e)) the first impression is that the

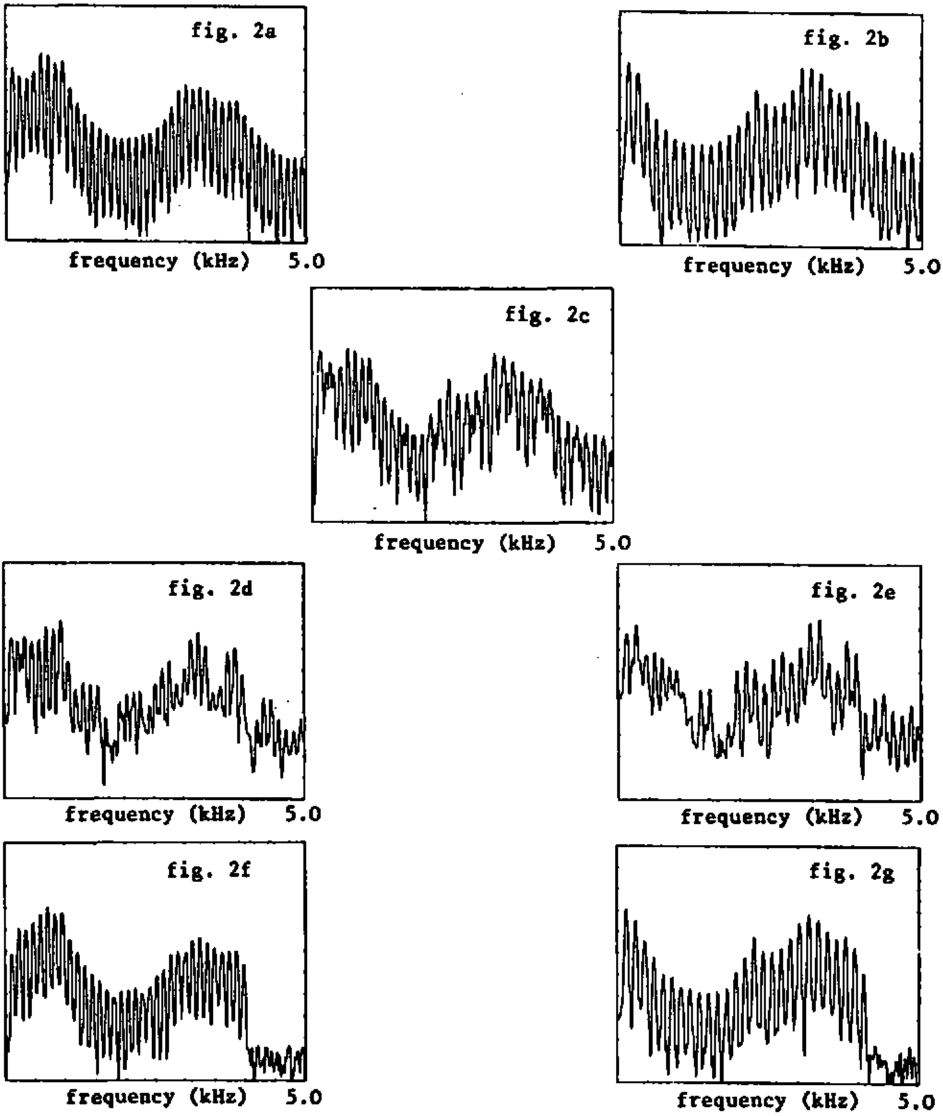

FIGURE 2. Log amplitude spectra of a) synthesised AH, f0 = 120 Hz b) synthesised EE, f0 = 151 Hz c) combined AH and EE d) recovered AH, cepstral filtering e) recovered EE, cepstral filtering f) recovered AH, harmonic selection g) recovered EE, harmonic selection

regularity which characterised the original separate spectra has been destroyed. However, comparison between the fundamental frequencies and formant frequencies of figs(2a,2b,2d,2e) reveals that the recovered spectrum of fig(2d) is similar to that of the AH, and that of fig(2e) is comparable to the EE. In fact, the recovered signals are perceived clearly as AH and EE. On examples containing naturally spoken harmonic speech the effect of the filtering is to reduce the unwanted voice to a hoarse whisper. The amount of attenuation is less than for synthesised sources, due to the reduced harmonicity in a natural voice.

HARMONIC SELECTION

Harmonic selection (4) is a form of 'harmonic vocoding', in which the amplitude spectrum is reduced to a table containing the centre frequency and amplitude of all harmonic peaks. Subsequent to a pitch determination, this information is used to resynthesise the constituent voices.

The centre frequency of resolved harmonic peaks and of significant shoulders on peaks due to unresolved harmonics is determined by successive differentiation of the amplitude spectrum. The amplitude of each feature is determined from a parabolic least-squares fit to the four samples nearest to its centre frequency. Features are subjected to four quality tests, which involve measures of the phase stability across a feature and of the proximity and relative amplitudes of neighbouring features. The quality of a feature is used as a weighting factor in subsequent operations.

The heart of the harmonic selection method is a pitch determination algorithm which is capable of extracting the pitches of simultaneous voices using the table of peaks. There are many pitch determination techniques (6) but in general they are designed for a single speaker in situations with a high signal to noise ratio. We have used a modified Schroeder histogram (4,7) but but have realised it in a log2 matrix (c.f. the spiral pitch processor (8)). The algorithm is constrained to a single octave pitch range from 100 to 200Hz. To determine the pitches of two voices the algorithm is used twice. When the first pitch has been calculated those features which are consistent with this harmonic series are flagged. The pitch determination is repeated using only unflagged peaks and those features consistent with the second pitch are flagged. It is possible for a peak to be flagged as belonging to both series.

The individual amplitude spectra of the two voices are constructed using the information in the peak tables. For each voice, those features which have been flagged as belonging to only that voice are used. A Hanning window lineshape (9) is used to create harmonics of the relevant centre frequencies and amplitudes. The central ten sample points of each harmonic are calculated and the phase at the peak centre is taken from the original FFT phase spectrum and imposed across the peak. In this way a fragmentary spectrum is created, with many 'missing' harmonics whose centre frequency can be calculated, but whose amplitude is unknown. The amplitudes of missing harmonics are linearly interpolated from nearest neighbours. Having constructed the two signal segments, they must each be attached to the appropriate output signal sequence. In allocating the segments, Parsons (4) invoked pitch continuity with previous segments. To date, our implementation is less sophisticated and simply sends the higher-pitched voice to one sequence and the lower-pitched voice to the other.

An example of harmonic selection separating the same two concurrent synthesised vowels previously discussed for cepstral filtering is shown in fig(2). The recovered spectra (figs(2f,2g)) are considerably more similar to the originals (figs(2a,2b)) than are the cepstrally filtered versions. Separation of naturally spoken harmonic sentences also produced good results. Subjectively the separation achieved by harmonic selection is superior to that of cepstral filtering. The recovered voices are intelligible, both for synthetic and natural speech, and in the latter case the voices are unmistakably human and the speakers are identifiable. Unlike the cepstral filtering which attenuates but does not completely remove the unwanted voice, harmonic selection produces a signal which contains only one source, since the synthesis creates only one harmonic series. The success of the process is largely dependent upon the efficiency of the pitch determination algorithm. Given two harmonic sources, errors are usually isolated and confined to the second pitch to be found in a segment.

PERCEPTUAL EXPERIMENTS

The algorithms were evaluated using a test involving the separation of concurrent vowels. Stimuli were created using a cascade formant synthesiser (10). Five vowels AH, EE, OO, OR and ER were synthesised on pitches of 120, 122, 124, 127, 135, 151, 171 and 191 Hz. The same intensity of harmonic excitation was used for all vowels, resulting in natural differences in overall amplitude between the vowels. The vowels were paired in all possible combinations, containing one vowel on 120 Hz and one chosen from a higher pitch. The vowel pairs were processed to enhance either the high- or low-pitched vowel by both processing methods. Synthesised stimuli were used to provide an invariant spectral envelope for a vowel, whilst allowing freedom in the choice of pitch. The sacrifice in using this method is to create stimuli which are more regular and overtly harmonic than naturally spoken vowels. The test results may therefore reflect an absolute performance level rather than an average performance for naturally produced tokens.

Four normal-hearing and four impaired listeners were tested. The impaired subjects had moderate to severe, symmetrical, sensorineural losses, making them candidates for future signal-processing aids. The identification task was run in two conditions. In the first condition, subjects were presented with a randomised sequence of unprocessed vowel-pairs and pairs processed by cepstral filtering or harmonic selection to enhance the low-pitched vowel. Subjects were asked to identify the low-pitched vowel in each pair. In the second condition the high-pitched vowels were enhanced and subjects were asked to identify the the high-pitched vowel. In order to observe an increase in performance level created by the processing it is necessary to establish a significant error rate for the unprocessed stimuli. This was achieved by using a stimulus duration of 100ms for the normal-hearing subjects and 280ms for the impaired subjects.

The test data for the normal and impaired subjects in the two conditions of the task are shown in figs(3a,3b,3c,3d). Each graph shows the percentage of errors

Figure 3a.

Figure 3b.

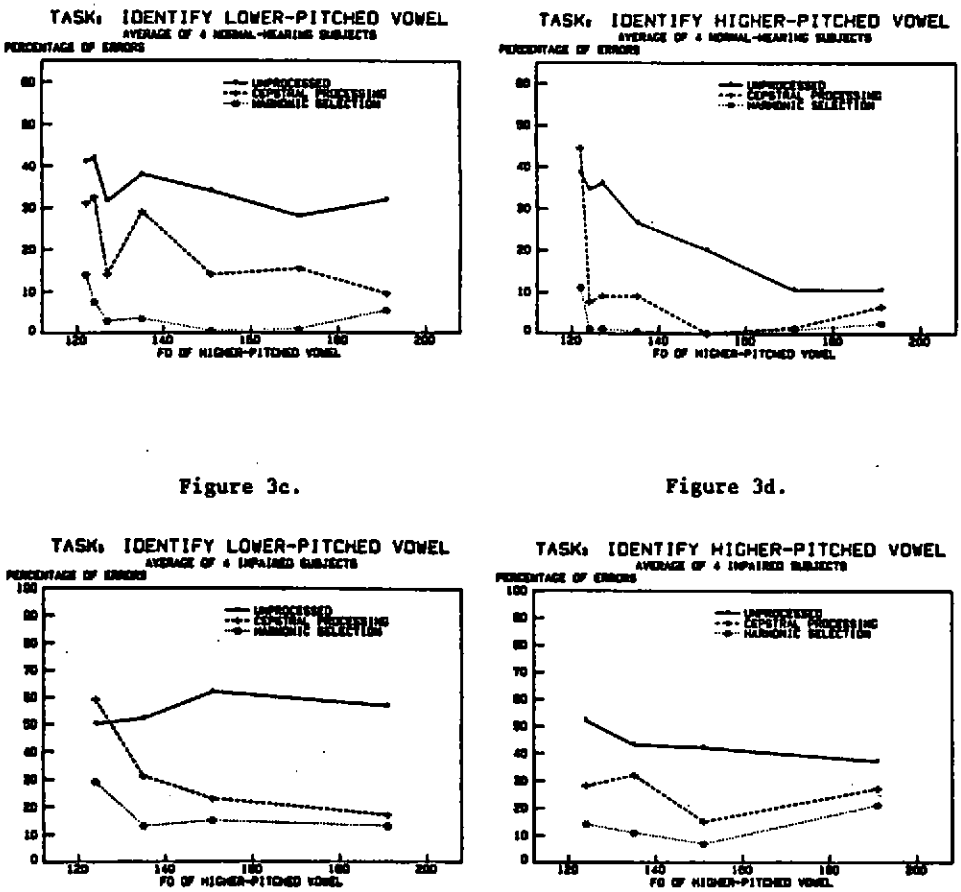

Figure 3c.

Figure 3d.

FIGURE 3. Results of the vowel separation test for a) and b) normal-hearing subjects c) and d) hearing-impaired subjects

made in the identification task, versus the fundamental frequency of the high-pitched vowel in the pair. Thus the x-axis represents increasing pitch separation between the vowels in the pair. Each graph shows three performance rates; for unprocessed vowel-pairs, for pairs processed by cepstral filtering and for pairs processed by harmonic selection. Results have been averaged over the four subjects within each group, but are representative of the data obtained from each subject. The major difference between subjects in a given group was in their absolute performance levels. For both normal and impaired listeners, the processing improved performance for a pitch separation of greater than 1/2-semitone (3%). The harmonic selection process gave a larger increase in performance than the cepstral filtering. In the case of the unprocessed pairs, the normal listeners reported that they were frequently able to hear both of the vowels in the pair and choose between them. The impaired listeners reported that they were unable to this.

CONCLUDING REMARKS

The results of the evaluation of the two voice separation algorithms demonstrate that both are capable of improving the performance of normal-hearing and hearing-impaired listeners in a voice separation task. The vowel separation test and subjective listening suggest that the harmonic selection method produces the better results. It is encouraging that this approach is more successful since it has greater potential for further development. The practical application of the algorithms is extremely limited at present. If they are to progress further, to natural speech signals, then at the very least a pitch tracking mechanism of the type used by (4) must be introduced. In the case of harmonic selection a voiced/unvoiced/silent categorisation is required for each of the competing voices.

REFERENCES

(1) M. Weintraub, "A theory and computational model of auditory monaural sound separation". Ph.D. dissertation. Dept. of Elec. Eng., Stanford University. (1985)
(2) B.C.J.Moore, "Frequency selectivity and temporal resolutionin normal and hearing-impaired listeners". British Journal of Audiology, 19, 189-201, (1985)
(3) W.A.Dreschler and R.Plomp, "Relation between psychophysical data and speech perception for hearing-impaired subjects I ". J.A.S.A. 68, 1608-1615, (1980)
(4) T.W.Parsons, "Separation of speech from interfering speech by means of harmonic selection". J.A.S.A. 60, 4, 911-918, (1976)
(5) L.R.Rabiner and B.Gold, "Theory and application of digital signal processing". Prentice-Hall. (1975)
(6) W.Hess, "Pitch determination of speech signals". Springer-Verlag. (1983)
(7) M.R.Schroeder, "Period histogram and product spectrum: new methods for fundamental-frequency measurement". J.A.S.A., 34, 829-834, (1968)
(8) R.Patterson, "Spiral detection of periodicity and the spiral form of musical scales". Psychology of Music, 14, 41-61, (1986)
(9) T.W.Parsons and M.R.Weiss, "Enhancing/Intelligibility of speech in noisy or multi-talker environments". RADC-TR-75-155 tech. report, Nicolet Scientific Corp. (1975)
(10) D.H.Klatt, "Software for a cascade/parallel formant synthesiser". J.A.S.A., 67, 971-995, (1980)