

SEPARATION OF SIMULTANEOUS VOICES

Richard J. Stubbs* and Quentin Summerfield*

ABSTRACT

The ideal speech-processing system, taking its input from a noisy environment, would be able to select a target voice whilst attenuating competing voices and other background noises. In exploring possible noise-reduction strategies, we have evaluated two algorithms designed to separate the voices of talkers who are speaking simultaneously, with periodic excitation at similar overall intensities. Both approaches are pitch-based and exploit the regularity in the harmonic structure of voiced speech. The first involves attenuating the periodic excitation of the competing voice via a cepstrum. The second method is derived from the procedure for Harmonic Selection (ref 1). Perceptual evaluation of the two processing methods, in tests involving the separation of simultaneous vowels, monotone sentences and naturally-intoned sentences, has demonstrated a significant increase in performance for normal-hearing subjects. Improvements in performance in tests involving subjects with sensorineural hearing-impairments suggest possible applications in future digital signal-processing hearing-aids.

INTRODUCTION

This paper describes the initial evaluation of two algorithms designed to separate the voiced speech of two talkers who are speaking simultaneously, at similar intensities. The algorithms exploit the harmonicity of voiced speech and require a difference in fundamental frequency between the voices to operate successfully. Prior to processing, the two voices are combined in a single channel. Thus the perceptual evaluations assess monaural segregation abilities.

We are primarily interested in voice-separation algorithms as examples of speech enhancement techniques which might form a part of future digital signal-processing hearing-aids. However, such techniques are equally applicable as pre-processing modules in speech-recognition or communication systems. The sound of competing talkers is a very common background noise, which is particularly detrimental to speech intelligibility, since the noise rejection task involves the segregation of signals with similar spectro-temporal properties. The ability of listeners with normal hearing to attend selectively to one of a set of competing voices is of interest in itself, but we feel that the exploration of voice-separation algorithms may illustrate general methods for extracting speech from complex, unpredictable interfering noises.

The two algorithms described here are of limited applicability and should not be construed as comprehensive voice-separation strategies, rather they might form a part of a segregation scheme. Since they only work on voiced speech, about 50% of natural speech is inaccessible to them. However, the voiced portions are the most intense parts of speech and they have a regular physical structure. Thus they are the obvious place to start

*

MRC Institute of Hearing Research, University of Nottingham,
University Park, Nottingham NG7 2RD, United Kingdom

voice separation.

PROCESSING

The two algorithms have a similar overall structure. A signal containing two voices is digitised at 10kHz and divided into overlapping 51.2ms hanning-windowed segments, which are processed sequentially. Each segment is Fourier transformed to produce an amplitude spectrum, in which harmonic structure is observed, and a corresponding phase spectrum. The amplitude spectrum, containing both voices, is processed to separate it into the two constituent spectra, which are individually combined with the original phase information and transformed back to the time domain.

The two methods of processing the amplitude spectrum are quite different. The first will be referred to as 'Cepstral Filtering'. This is a noise reduction technique. It assumes that the noise interfering with the target voice is harmonic. The combined amplitude spectrum is transformed to produce a cepstrum. If two harmonic sources which have different fundamental frequencies are present in the signal, two pitch peaks are observed in the cepstrum. By zeroing the pitch peak of one of the voices it is possible to attenuate that voice.

The second method, shown in Figure 1, is a reworking of the technique of 'Harmonic Selection' (ref 1). This is a signal extraction technique. It assumes that the desired target voice is harmonic. The centre frequency and amplitude of all harmonic peaks, and shoulders on peaks, in

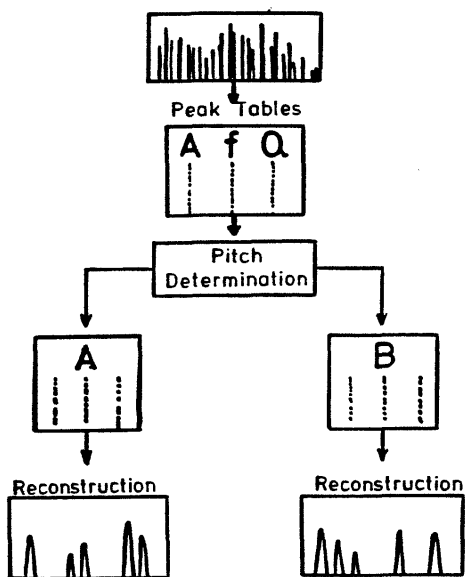


Figure 1. Harmonic Selection

the combined amplitude spectrum are determined. The peaks are assigned a quality weight and entered into a table of peaks. This information is submitted to a pitch determination algorithm which calculates the two best fitting harmonic series which underly the peaks and shoulders. Those harmonics which can be assigned exclusively to pitch A (i.e. Voice A) are grouped together and similarly those which belong exclusively to pitch B (i.e. Voice B). This fragmentary evidence is used to reconstruct the voices. Using a hanning lineshape, the known harmonics of a voice are inserted into a blank amplitude spectrum. There will be many missing harmonics whose centre frequency can be calculated, but whose amplitude is unknown. These amplitudes are interpolated from nearest known neighbours and the reconstruction is completed.

We have evaluated the two algorithms using three tests, designed to be of increasing difficulty both for the algorithms and for listeners. To date, we have run hearing-impaired subjects on the first test only. This test, which has been reported previously (ref 2), involved the separation of pairs of synthetic vowels with static fundamental frequencies. Four normal-hearing and four hearing-impaired subjects were asked to identify vowels which had been masked by the addition of another vowel. The

performance of both subject groups was increased by between approximately 10% to 50% when the target vowel had been enhanced by processing and the pitch difference between target and masker was greater than 5Hz. Processing by Harmonic selection gave significantly greater improvement than Cepstral Filtering. These results with perfectly harmonic stimuli confirm the basic ability of the algorithms to separate the combined spectral amplitude information on the basis of fundamental frequency.

The second and third tests involved the separation of voiced sentences spoken by human, rather than synthetic, talkers. An example is: "We are mining a mineral in our area". For the second test, twenty-eight such sentences were uttered on a monotone by one male speaker on a fundamental of approximately 120Hz and by another on a fundamental of approximately 155Hz. Sentences were normalised to the same RMS level, matched in duration and mixed in pairs. Each pair consisted of two different sentences, spoken by the different talkers. These mixed sentences were processed by each of the two algorithms to enhance either the lower or higher-pitched talker. Eight normal-hearing subjects were presented with a randomised sequence consisting of mixed sentences, isolated sentences spoken by one of the two talkers, and mixed sentences processed to enhance that same talker. Subjects (who were familiar with the two talkers) were given an example of an isolated sentence spoken by the target talker and were told to attend to his voice throughout the test and write down what he said on each trial.

Performance was scored as the number of key words correctly identified in each target sentence. The results are illustrated in the left-hand bar graph in Figure 2. When presented with isolated sentences, subjects identified 98% of the key words correctly, whereas in unprocessed pairs of

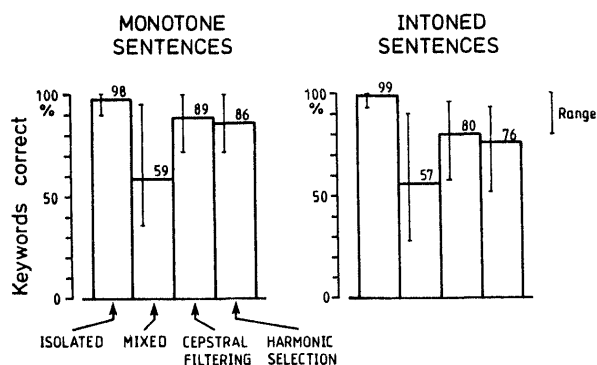


Figure 2. Results of Test 2 (left) and Test 3 (right).

sentences they identified only 59% of the keywords correctly. Both processing methods gave significant increases in intelligibility. In this test, as in the first, the fundamentals of the voices were relatively constant. However, the sentences were produced by humans rather than by a synthesiser and the spectrum envelope of each speaker changed over time. The results demonstrate that the algorithms do not require perfectly harmonic input signals and that they preserve time-varying spectral information.

The third test was similar to the second, except that the talkers spoke with natural intonation. Natural intonation has two consequences for the segregation algorithms. Firstly, the pitch is not constant throughout the duration of a processing segment, causing a blurring of harmonics in the spectrum, particularly at higher frequencies. This makes it harder to determine the pitch and to manipulate the spectrum on the basis of pitch. Secondly, the pitch tracks of the two voices may cross or coincide. It is no longer adequate to process on the basis of a higher or lower pitched voice; rather the processing requires a pitch tracking mechanism which

can lock on to and follow the pitch of the target voice. Parsons (ref 1) suggested using a linear predictive filter to do this. We have tried several orders of filter and have treated the pitch of a voice as a two dimensional function of frequency and 'pitch strength' but have found this inadequate to cope with many situations that arise when naturally-intoned sentences are combined. It is probable that additional constraints will have to be used to track the target voice, most likely spectrally based. However, in order to evaluate the segregation algorithms, rather than the control mechanisms, in preparing stimuli for the third test we steered the processing 'correctly' through pitch crossings and coincidences.

The same eight subjects were tested again. The results are illustrated in the right-hand bar graph in Figure 2. When presented with isolated sentences, subjects identified 99% of the target words correctly. With unprocessed pairs the performance was 57% correct. These performance levels are almost identical to those found with the monotone sentences, which suggests that the inclusion of natural intonation has not made the sentences perceptually more, or less, separable. Processing improved performance significantly, although the improvements were smaller than those achieved with the monotone sentences.

These results allow four conclusions:

1. Cepstral Filtering and Harmonic Selection are capable of separating simultaneous voiced speech sources by accessing them via their harmonic structure.
2. Voices separated in these ways are more intelligible than they are prior to processing.
3. With perfectly harmonic synthetic test stimuli, Harmonic Selection gives greater enhancement than Cepstral Filtering. With voiced speech produced by real talkers, the pattern is reversed (though the advantage for Cepstral Filtering was not statistically significant). This result may have occurred because Harmonic Selection, being the more analytic procedure, is capable of producing more catastrophic errors. Harmonic Selection has the additional disadvantage of being more expensive computationally than Cepstral Filtering. These disadvantages may be balanced by two advantages. Harmonic Selection achieves enhancement by identifying and grouping the components of the target voice, whereas Cepstral Filtering filters out the interfering noise. This difference has two consequences. First, Harmonic Selection should be able to enhance voiced speech corrupted by a wider range of background noises than Cepstral Filtering. Second, a signal reconstructed by Harmonic Selection sounds like a single voice, while a signal reconstructed by Cepstral Filtering contains vestiges of the interfering voice.
4. Predictors based only on the physical shape of the fundamental-frequency contour of a voiced sentence do not successfully track that contour in the presence of a competing contour. Overcoming this problem is the key to the successful application of algorithms such as these.

REFERENCES

1. T W Parsons, JASA, vol.60, 911-918 (1976)
2. R J Stubbs and Q Summerfield, Proc. IOA: Speech and Hearing, vol.8, 7, 119-126 (1986)