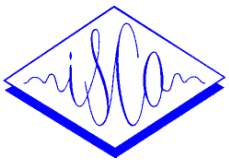

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

A SYSTEM FOR SPEECH SEPARATION

A.Shamsoddini and P.N.Denbigh
email: a.shamsoddini@sussex.ac.uk
p.n.denbigh@sussex.ac.uk
School of Engineering
University of Sussex
Brighton BN1 9QT
U.K.

ABSTRACT

A system has been developed that successfully separates a target voice from an overlapping voice. The algorithm is based mainly on monaural processing, exploiting the harmonicity of voiced speech, but also utilises some binaural information to initially allocate each separated sound to the appropriate speaker. Emphasis has been put on monaural processing because of the sensitivity of binaural cues to room reverberation.

1. INTRODUCTION

Normal listeners are able to "focus" onto a desired voice in presence of other interference voices, an ability referred to as the "cocktail party effect". There are cases however when this ability of sound separation is degraded or lost. A person with a hearing impairment of cochlear origin finds it difficult to concentrate on the wanted voice in the presence of interference. Interfering speech also creates difficulties for speech recognition systems. A successful speech separation system has potential for use as an intelligent hearing aid, as a pre-processor for a speech recognition system, or as an enhancement unit in a communication system.

Many researchers have tried to make sound separation systems using only monaural cues, based upon extracting the pitch harmonics of the target voice [e.g. 1,2]. The main problems have been

a. to determine the pitch frequencies when there is more than one voice.
b. to know which pitch belongs to the target voice.

Another major approach to the sound separation problem has been based purely on directional cues. These systems utilise the time and/or intensity differences from the signals received by two or more microphones to extract the wanted voice. The performance of these systems is however greatly degraded by multipath echoes in non-anechoic rooms. For example, the response of a 6m by 4m office, measured by recording the impulsive sound of a child's cap gun at a distance of 2m, is shown in Fig.1. It shows that the majority of the incident energy does not come via a direct path.

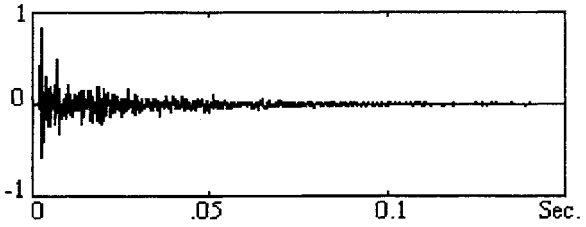

Fig.1 Impulse response of 6m by 4m office.

Denbigh and Zhao [3] devised a hybrid system that was a combination of monaural and directional systems. Pitch was first acquired at the *onset* of a voiced sound by applying Hermes' subharmonic summation method of pitch extraction [4] to the *change* in the spectrum between successive 40ms time frames 20ms apart. This pitch was then *tracked* by subsequently constraining the spectrum entering the subharmonic summation pitch extractor to those frequency regions close to the harmonics of the previously measured pitch. The averaged time differences of these harmonics between two separated microphones were then used to help an extracted pitch to be attributed to the appropriate speaker on the basis of direction. Luo and Denbigh [5] and Denbigh and Luo [6] added some extra features to the system. The main change was to track the harmonics of each voice using a "two dimensional harmonic sieve". As with Denbigh and Zhao's system this allowed spectral lines to pass to a pitch estimator only if they were close in frequency to harmonics of the previously estimated pitch. However it *also* demanded that they were close in *amplitude* to pitch harmonics of the previous frame. Thus the two dimensional harmonic sieve was based on both common frequency *and* common amplitude modulation cues. For the case of minor reverberation the system produced good intelligibility scores even when the target voice was 12dB weaker than an interfering voice [5,6]. However, for levels of reverberation encountered in many real-life situations, its performance was considerably degraded.

One reason for this is that, besides degrading the directional cue, reverberation also degrades spectral peaks. For example the frequency response corresponding to the impulse response of Fig.1 is shown in Fig.2 and it will be clear that its closely spaced peaks and troughs can have a significant effect on the magnitude and frequencies of spectral peaks caused by pitch harmonics.

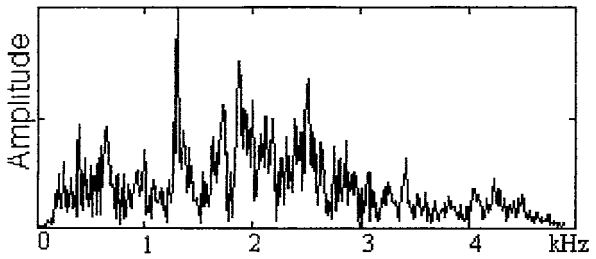

Fig.2 Frequency response corresponding to Fig.1

The work to be described is a continuation of that described above and is aimed

- at being more robust to reverberation
- at being less demanding computationally such that the algorithm is capable of being realised in a real time system.

2. ALGORITHM

The separation system is designed for the case of a target voice and one interfering voice. At low frequencies the separation algorithm is based on a combination of harmonic enhancement of the wanted voice and harmonic cancellation of the unwanted voice. The pitches of both voices are tracked. When the target speech is voiced it is enhanced by passing only the harmonics of its pitch. When the target speech is unvoiced but the interfering speech is voiced, enhancement is achieved by eliminating the pitch harmonics of the interfering speech. When both speech signals are unvoiced, separation is achieved using directional cues.

In the high frequency region of the spectrum a directional technique of cancelling interference is used.

There are two microphones 25cm apart. The sampling frequency is 10kHz and processing is based on 40ms frames with 50% of overlap. Each frame of the output of each microphone is weighted with a Kaiser-Bessel window and then padded with zeros to occupy 102.4ms. This produces FFT coefficients just less than 10Hz apart. A simplified block diagram of the system is shown in Fig.3.

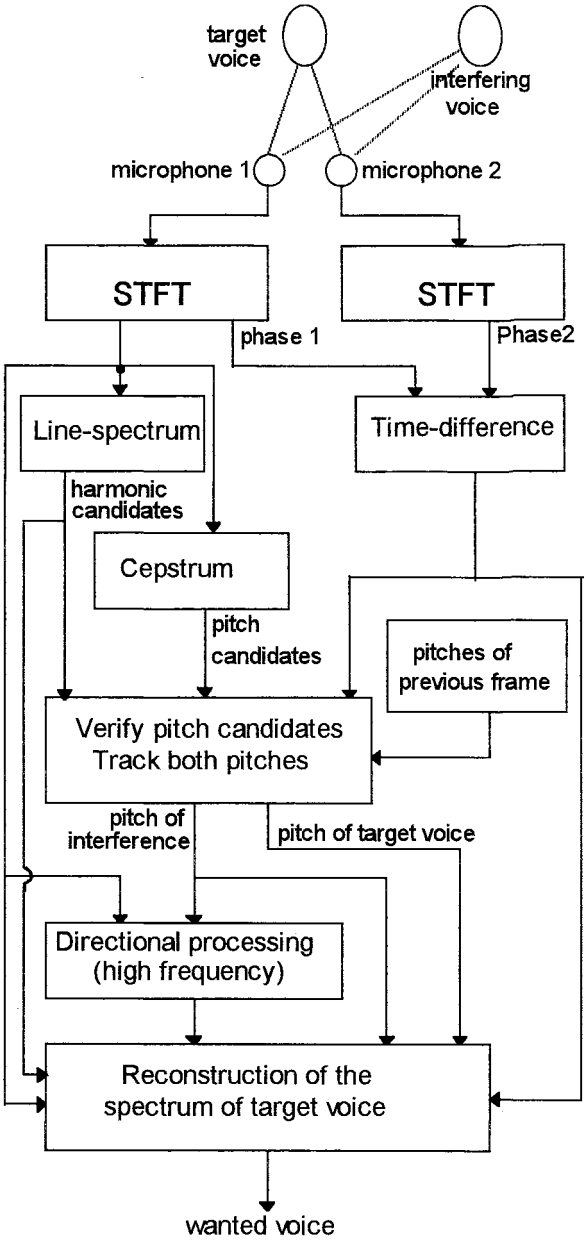

Fig.3 Simplified block diagram of the sound separation system.

2.1. Pitch Extraction and Tracking

Compared with earlier systems developed by this group, computational efficiency has been improved by replacing Hermes' subharmonic summation method of pitch extraction with Noll's cepstral method [7]. A difficulty to be overcome with two voices however is the presence of spurious, degraded, and displaced peaks in the cepstrum created by the effect of having more than one signal in the input. The problem is overcome by a system of acquiring, verifying, and then tracking the true pitch peaks in the cepstrum. With reference to Fig.3 the strategies include the following.

- in parallel with the cepstral analysis a conventional spectrum is produced which is then simplified into a set of spectral lines at the spectral peaks and shoulders. Because these frequencies can be slightly in error, extra lines are added on either side to create what is termed the "line spectrum".
- the period of a peak in the cepstrum is considered to be a possible pitch period if the amplitude of that peak is bigger than a predetermined threshold. The reciprocal of such a pitch period is termed a *pitch candidate*.
- a pitch candidate is accepted as a *true* pitch if a certain number of its harmonics are detected in the line-spectrum (the exact number required depends on the pitch frequency and changes linearly between 11 and 4 as the pitch changes between 70 and 400Hz).
- after the *initial* acquisition of a pitch in this way, usually made possible by the momentary dominance of that voice, the cepstral peak is *tracked* in subsequent frames. This is done by giving preference to cepstral peaks in the immediate vicinity of the previously determined pitch peak, and is achieved by lowering the amplitude threshold in that cepstral region.
- the target speaker is always taken to be directly broadside to two displaced microphones. An *initial* attribution of a pitch peak to this speaker is made by demanding that the averaged time difference of the associated harmonics should be less than 150μs (corresponding to less than 12 degrees). Otherwise it is attributed to the interfering speaker. In subsequent frames a dynamic demand on direction is imposed as part of the tracking procedure. This allows a pitch to be allocated to a voice by looking at its average time difference and allocation in previous frames, as well as its average time difference in the current frame.

By the above procedure, it is expected that the system,

a. determines when the voicing of each speech begins or ends.
b. stays locked to each pitch so long as voicing continues, even at times when their cepstral peaks are degraded.
c. attributes the pitches of voiced segments to the correct one of the two speakers.

2.2. Reconstruction of target voice

For reconstruction purpose the spectrum is considered as divided into a low frequency region and a high frequency region. The high frequency range is treated differently because the harmonic structure tends to be more diffuse in this region, especially in the presence of interference or reverberation.

When the target speech is voiced but the interfering voice is unvoiced, the components in the line spectrum that lie below 3kHz and correspond to pitch harmonics are passed to be used for reconstruction. An exception is when *both* speech signals are voiced. When this happens, any harmonics of the target voice that are *shared* by target and interfering voice are rejected unless it is clear that they belong predominantly to the target voice. This decision is made on the basis of direction from which these harmonic components arrive, or on the basis that these harmonics are similar in amplitude (common amplitude modulation) to harmonics previously associated with the target voice in the previous frame.

When the target speech is unvoiced but the interfering speech is voiced, the harmonics of the interfering speech below 3kHz are removed from the line spectrum before the line spectrum is used for reconstruction.

When both speech signals are unvoiced, coefficients in the line spectrum below 2kHz are passed for reconstruction if their time difference between the two microphones corresponds with the target direction.

A different approach is used for selecting high frequency spectral coefficients (i.e. above 3kHz in the case of voicing, or above 2kHz when both are unvoiced). This is done using a technique of interference cancellation that is based on directional information. This directional information is obtained from the pitch tracking section, based on the frame-average of the harmonic-averaged time differences of the unwanted voice, for the case when the unwanted voice is voiced. By averaging over an adequate number of frames, this parameter is an indication of the delay between the unwanted voice components received by left and right microphones. With the assumption that the target speaker is directly in front of the microphones, we first subtract the output spectra of the two microphone outputs to *reject* components of the *wanted* voice. The result of this subtraction then is used along with the information of the above-mentioned time delay to reconstruct the complete spectrum of *unwanted* voice. The calculated spectrum of unwanted voice is then subtracted from the mixed spectrum to give the spectrum of the *wanted* voice.

Although the performance of this method of directional processing is degraded by reverberation, it is only used where and when harmonic processing is not possible (above 3kHz, and at unvoiced instants) and improves the performance of the system.

3. RESULTS

As stated earlier, the system can produce very effective voice separation even at low ratios of signal to

interference (as low as -12dB). As an example of its performance in the absence of reverberation, Fig.4(a) shows the waveform of the utterance "he's wiping the table", and Fig.4(b) shows it contaminated by an interfering voice 6dB stronger.

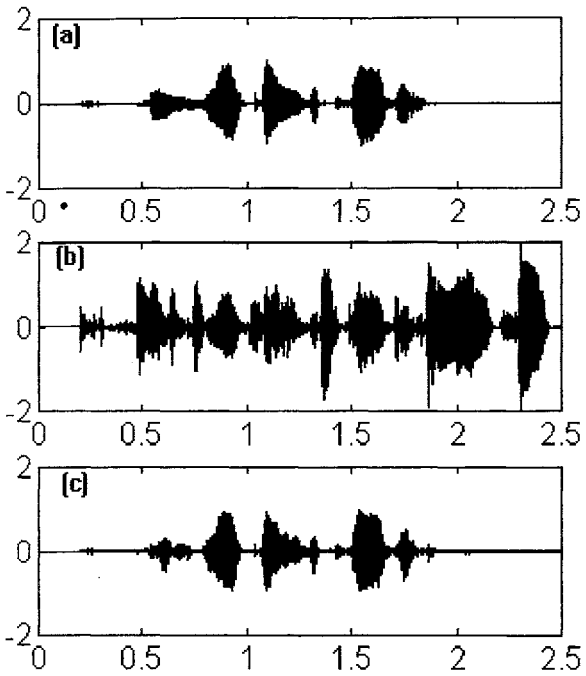

Fig.4 Time waveforms of single, overlapping, and separated speech.

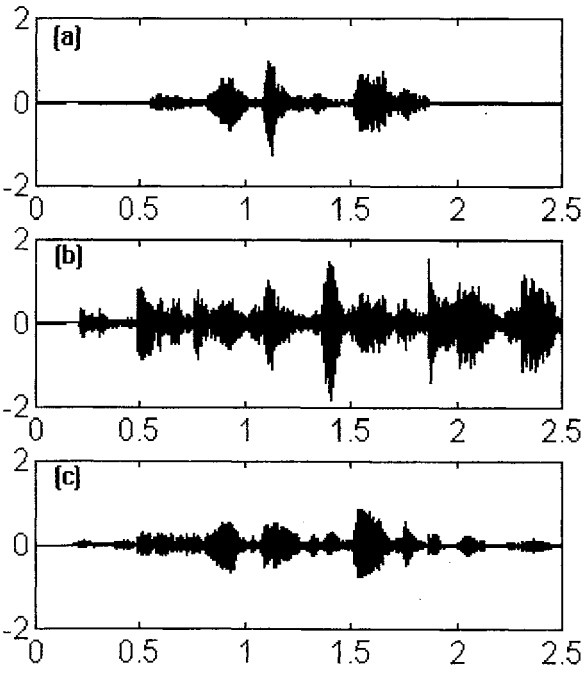

Fig.5. Time waveforms of single, overlapping, and separated speech in a reverberant situation.

Fig.4(c) shows the waveform extracted by the system from the two overlapping speech signals and it is seen to closely resemble that of Fig.4(a).

The same three cases are illustrated in Fig.5 for the case of moderate reverberation. The reverberation was simulated but was comparable to that which produced Fig.1.

Subjective tests show that the outputs of Fig.4(c) and Fig.5(c) are both highly intelligible; also that there is a great gain in the ease of listening.

4. DISCUSSION

The greatest problem encountered is that of making the system robust against reverberation, since this greatly degrades the directional information. It is because of reverberation that separation is based primarily on pitch and harmonic analysis rather than on direction. Although the performance of the proposed system degrades with reverberation, it is more robust than systems which are solely directional. The system is computationally undemanding and could be realised in real-time using available signal processing technology.

5. REFERENCES

[1] Parsons, T.W., "Separation of speech from interfering speech by means of harmonic selection", J. Acoust. Soc. Am., No. 60 (1976), 911-918

[2] Stubbs, R.J. and Summerfield, Q., "Algorithms for separating the speech of interfering talkers: Evaluation with voiced sentence, and normal-hearing and hearing-impaired listeners", J. Acoust. Soc. Am. vol. 87 (1990), No. 1, 359-372

[3] Denbigh, P.N. and Zhao, J., "Pitch extraction and separation of overlapping speech", Speech Communication , No. 11 (1992), 119-125

[4] Hermes, D.J., "Measurement of pitch by subharmonic summation", J. of Acoust. Soc. Am., No. 83, vol. 1, (1988), 257-264

[5] Luo, H.Y. and Denbigh, P.N., "A speech separation system that is robust to reverberation", IEEE Intl. Symposium on Speech, Image Processing and Neural Networks, 13-16 April 1994, Hong Kong, 339-342

[6] Denbigh, P.N. and Luo, H.Y., "An algorithm for separating overlapping voices", IEE Colloquium for Speech Processing and their Application., Digest No. 1994/138, pp. 9/1-9/6.

[7] Noll, A.M., "Cepstrum pitch determination ", Journal of Acoust. Soc. Am., No. 41 (1967), 293-304