AUDITORY GROUPING AND ATTENTION TO SPEECH

C.J.Darwin Experimental Psychology, University of Sussex, Brighton BN1 9QG.

1. INTRODUCTION

The human speech recognition system is superior to machine recognition in many ways, but one of the most dramatic is in its resistance to additive "noise" [1]. The human listener is able to use a variety of types of information to help segregate the sounds from different sound sources. The pioneering work on the basic phenomena of human sound segregation was carried out by Al Bregman, and is reported in his book "Auditory Scene Analysis" [2] together with a theoretical analysis of the problem and types of solution. Bregman's own work has been mainly concerned with simple musical sounds; our own work [3, 4] has concentrated on exploring the relevance of auditory scene analysis to speech perception. This paper will briefly review some of the results from this work that may have implications for the machine recognition of speech.

The need for some segregation of one sound from another in the process of recognition is evident from a number of observations such as:

- Stored information about sounds (e.g. acoustic/phonetic relations) probably concerns a single source whereas the raw spectral data presented to the brain is a mixture. Features necessary for recognition (e.g. silence in stops, can only be defined in a source-specific way [5]).
- We perceive the individual sounds of sound mixtures with (broadly) their appropriate timbres, pitches and locations

The interesting issues are:

- To what extent this segregation relies on specific knowledge of different sounds or classes of sound, rather than low-level general cues such as harmonicity and onset-time?
- What are the specific psychological mechanisms that give rise to segregation, and what are their physiological bases?
- Can a general statistical procedure such as ICA, wedded to an appropriate auditory front-end explain the human data?

1.1 The need for low-level grouping of speech

It has been claimed by Remez and others [6] that speech is immune from low-level auditory grouping. A primary argument for this claim is that phonetic coherence of sine-wave speech cannot be explained on purely auditory principles. Although this argument has some merit, it has a number of shortcomings. First, sine-wave speech is an abstract representation of speech which makes formant frequencies explicit. We have shown that auditory grouping is a useful concept prior to the level at which the first formant frequency is identified [7]: a harmonic can be separated from a vowel by virtue of their different onset times even though such segregation leads to a vowel lacking a harmonic. Here low-level grouping is constraining the alternatives that top-down or speech-specific mechanisms can consider. Second, although the sine-wave speech of a single talker is intelligible, a mixture of two sine-wave speech sentences is much less intelligible than is a mixture of two natural voices [8]. This result is surprising since one might on the basis of masking expect the spectrally sparser sine-wave speech to maintain its intelligibility in a mixture better than natural speech. Third, the auditory grouping cues of common onset and continuity do provide some information about source assignment for sine-wave speech [8]. The obvious property that is absent in sine-wave speech is harmonic structure, and it is likely that this property is one that makes a substantial contribution to the human perceptual system's ability to separate two simultaneous voices.

2. USING Fo DIFFERENCES

One of the longest-established and most effective experimental manipulations for helping listeners to separate two voices is a difference in fundamental frequency (Fo).

2.1 Double vowels and continuous speech

Two vowels with a difference in Fo are more identifiable than two on the same Fo [9], but this improvement asymptotes by about 1-semitone difference in Fo and is probably due to beating between the low-numbered harmonics [10]. It is unlikely that double vowels provide a good model for the mechanisms used in normal speech for segregating the speech of two talkers, since resynthesizing sentences on increasingly different monotones gives an improvement in performance that extends out well beyond a semitone Fo difference [11]. This improvement can be very marked, particularly if sentences are chosen with few stops or fricatives so that onset cues to segregation are minimised [12]. There appear to be two different mechanisms responsible for this improvement which have been revealed using "chimeric" vowels and sentences [10] that have different Fos in their low- and high-frequency regions (Fig 1).

Complementary pairs of these sentences maintain a local difference in Fo that is similar to normal sentences but globally the difference in Fo provides inappropriate grouping. Using such chimeric (or Fo-swapped) sentences, and also semi-chimeric ones where one of the frequency regions has the same Fo in both sentences, we have been able to show that for small differences in Fo (=< 2 semitones) the improvement in intelligibility is due to local effects of Fo in the low frequency region, but not in the high frequency region, whereas for larger differences in Fo (=< 5 semitones) global (ie across-formant grouping) processes become important. So small Fo differences make it easier, say, to identify what the first formant of each talker is, whereas larger Fo differences also enable listeners to correctly group together the formants of a single voice. This difference between the effectiveness of Fo cues in the low and high frequency regions sits well with the psychoacoustic findings that the pitch sensation from low-numbered harmonics is much stronger than that from unresolved, high-numbered harmonics. It is when Fo differences are large that the weaker pitch information from the higher harmonics can successfully achieve differential grouping of the higher formants.

2.1 Frequency Modulation

Although putting a vibrato-like frequency-modulation onto a vowel or instrumental sound gives the sound greater prominence [13] and coherence [14, 15] than an unmodulated sound, a surprising but consistent result is that human listeners appear not to use such frequency modulation as a segregation cue. If two vowels differ in Fo, then giving them different frequency modulations does not increase their perceptual separation [13, 16] – a result which mirrors a remarkable basic psychoacoustic limitation in our ability to detect incoherent FM for inharmonic sounds [17]. These results illustrate the fact that the human auditory system does not always use information that is readily apparent to the eye in conventional or even "auditory" spectrograms. One possible reason for the auditory system's inability to perform segregation by frequency modulation is that coherent FM generally only appears on harmonics, which are strongly grouped anyway purely by virtue of their harmonicity. It is also possible that experiments using larger, slower frequency excursions might show some effect of segregation by incoherent modulation [18]. It would be interesting to know whether statistical learning algorithms can learn to exploit differential FM for the separation of two voices.

3. ONSET-TIME DIFFERENCES

A difference in onset-time between one component of a sound and the remainder is a very powerful cue for sound segregation. A single harmonic can be segregated from a vowel with consequent changes to the vowel quality [7], or, if the harmonic is slightly mistuned, to the vowel's pitch [19].

The use of onset-time differences in sound segregation is closely linked to Bregman's "Old+New Heuristic" whereby we try to maintain continuity of an initial sound when a new sound starts. What is then heard as the new sound is the total sound that is now present *minus* the old sound. Complementing the additivity of sound mixtures, this Old+New heuristic is subtractive within a frequency channel both for the loudness of noise bursts [[20, 21] and for vowel quality with subtracted harmonics [22]. Although the precise metric used for such subtraction is not clear, what is clear is that a single frequency channel is not exclusively allocated to one source or another. A minimalist demonstration of non-exclusive allocation can be made with a single frequency. Play the same sine-wave at equal amplitude to each ear, and then put a pulsed increment just on the left ear. You here a continuous steady tone in middle of the head, with additional pulsing tone at the same frequency on the left ear. Incidentally this demonstration argues against Kubovy's [23] notion that for two sounds to be heard they must differ in frequency (an "indispensible attribute" of a sound source).

A surprising feature of these results is that different values of onset time are needed to secure segregation as measured by vowel quality or by pitch. The pitch change requires a much larger onset-time than does the change in vowel quality [19]. By contrast, more mistuning of a harmonic is needed to remove it from the calculation of vowel quality than from the calculation of pitch. These observations indicate that perceptual grouping is NOT carried out in an all-or-none way prior to, and independently of, the perceptual classification task. Segregation algorithms which reconstruct separate sound sources are not therefore good models of human source segregation, since a particular frequency component may be regarded by the brain as part of a particular source for the purpose of calculating the sound's pitch, but not for the purpose of calculating its timbre.

4. LOCALISATION

Although localisation cues have been used quite extensively for the machine segregation of different talkers [24] an unexpected and important limitation on their use by human listeners has recently come to light. It is well-established that interaural time differences in the lower-frequency components are the dominant localisation cue for a complex sound such as speech [25]. However, such interaural time differences are remarkably ineffective at segregating simultaneous sounds. Culling & Summerfield [26] presented listeners with four narrow formant-like noise bands. If these noise-band pairs are led to opposite ears (eg bands 1&4 give /i/ and 2&3 give /a/), listeners have no difficult in hearing, say /i/ on their left ear . However, when Culling & Summerfield made their lateralisation manipulation with interaural-time differences rather than by leading the bands to different ears, their listeners were quite unable to do the task. We have then the apparently paradoxical result that the cue that is dominant for the localisation of complex sounds, is quite impotent for the grouping of simultaneous sounds. It is hard then to propose that interaural time differences should form the basis of an algorithm for sound segregation.

This unexpected result also seems to go against one's introspective experience of attending to one sound source rather than another. I certainly have the imporesssion that I am directing my attention spatially. Is this simply an illusion (though one supported by experimental evidence [27, 28]) or is the relationship between spatial attention and auditory localisation more complex than implied by the simple notion of attending to frequency channels that share the same inter-aural time difference?

Evidence supports the more complex relationship. The basic idea is that auditory grouping (based at least on primitive cues such as harmonicity and onset time) occurs prior to the localisation of complex sounds. According to this idea, which was initially proposed by Woods and Colburn [29], the interaural time (and intensity) differences (ITDs) of individual frequency channels are computed independently in parallel with a separate process which assigns these frequency channels to separate sound sources. The two types of information are then brought together so that the localisation of an auditory object can be constructed from the ITDs of the frequency components that make up that auditory object. Direct evidence for this ordering of events comes from experiments on the localisation of complex tones with abnormally large (1.5 ms) ITDs [30]. When all the components of such a tone have the same large ITD, the whole tone is heard on the side of the leading ear. However, if one of the frequency components is segregated from the complex by virtue of its onset time or mistuning, then it is localised separately on the opposite side of the head (the side that it would have been localised on – thanks to phase ambiguity - had it been the only component present).

Using natural speech sounds, resynthesised on various Fos, we [31] have found that natural sounds that only differ in ITD can be selectively attended very easily. For example, if two simultaneous monosyllabic words ("bead" & "globe") are given ITDs of ±90 μs respectively, they easily segregate into two spatially distinct auditory objects which can be readily attended to. For natural speech there are many cues (eg harmonicity, onset-time differences) which can help the auditory system to allocate individual frequency channels to the two different sound sources. What is more surprising is that the impression of two separate objects survives when the two words are resynthesised (using Praat's PSOLA) on the same Fo. For simpler sounds, such as steady vowels, the impression of two distinct sources with separate locations is destroyed by a common Fo. We are presently carrying out experiments to discover what cues are present in the natural speech which allow the segregation of two separately-located sound sources. It seems likely that higher-level constraints such as the particular formant tracks used are important in this ability.

5. ATTENTION AND GROUPING

There is more to attending to the voice of a particular talker than performing short-term segregation of frequency channels into sound sources. A particular sound source must be tracked across time. What cues do we use to achieve this? It might be the case that attention to speech is directed entirely spatially, so that we continue to listen to a particular spatial direction even if voice characteristics at that location change. Such a simple model for the direction of attention to speech is unlikely to be correct. We have demonstrated [32] that pitch-contour and vocal-tract-size cues can override spatial cues in determining which target word belongs in an attended sentence; a conclusion that we have confirmed using a real-time shadowing task. Listeners rapidly switch their attention when the attended voice switches to the opposite ear. Moreover, prosodic and vocal-tract-size cues are more resilient to degradation from reverberation than are ITDs [33].

The attraction of low-level grouping mechanisms is that they are pre-attentive. Low-level mechanisms help to sort out the combinations of frequency components that make up separate sound sources, allowing complex sounds to be localised and attended to. But the assumption of pre-attentiveness has recently been challenged by Bob Carlyon and his colleagues [34]. When an appropriate repeating, galloping rhythm (low-high-low-silence) is listened to for a few seconds, the single sound source is replaced by two sources, one at each pitch and the galloping rhythm is lost. This build-up of auditory streaming has been known for a long time and is well-established. However, the surprising result that Carlyon and his colleagues found was that attention to the sound is necessary for this build-up of streaming to occur. The time listeners spend attending to another sound in the opposite ear does not contribute to the build-up of segregation.

6. SUMMARY

This paper has briefly reviewed work on the perceptual segregation by human listeners of speech, emphasising findings which are surprising or not consonant with some particular algorithmic approach. Barker & Cooke's demonstration of listeners' very poor performance on two simultaneous sine-wave speech sentences compared with their performance on natural sentences is strong evidence that the segregation of speech cannot simply be accomplished by the type of information that is traditionally available for word recognition. The fine spectral detail available to the huiman listener, although of little value for recognition of a single talker, becomes of paramount importance when segregation of two talkers has to be achieved.

More work is needed in order to make clearer the role of higher-level constraints, such as speech-specific constraints on prosody and formant movement, and individual talker characteristics in allowing the segregation of localisable sources that can be tracked over time. The basic question of how attention interacts with source segregation is one that is also little understood.

REFERENCES

[1] R.P. Lippmann. Speech recognition by machines and humans. *Speech Communication, 22, pp 1-15,* 1997

[2] A.S. Bregman. *Auditory Scene Analysis: the perceptual organisation of sound.* 1990 Bradford Books, MIT Press. Cambridge, Mass.

[3] C.J. Darwin and R.P. Carlyon, Auditory grouping in *The handbook of perception and cognition, Volume 6, Hearing.* B. C. J. Moore(ed.), 1995, Academic. London.

[4] C.J. Darwin. Auditory grouping. *Trends in Cognitive Science, 1, pp 327-333,* 1997

[5] C.J. Darwin and C.E. Bethell-Fox. Pitch continuity and speech source attribution. *Journal of Experimental Psychology: Human Perception and Performance, 3, pp 665-672,* 1977

[6] R.E. Remez, P.E. Rubin, S.M. Berns, J.S. Pardo and J.M. Lang. On the perceptual organization of speech. *Psychological Review, 101, pp 129-156,* 1994

[7] C.J. Darwin. Perceiving vowels in the presence of another sound: constraints on formant perception. *Journal of the Acoustical Society of America, 76, pp 1636-1647.,* 1984

[8] J. Barker and M. Cooke. Is the sine-wave speech cocktail party worth attending? *Speech Communication, 27, pp* 1999

[9] M.T. Scheffers. *Sifting vowels: Auditory pitch analysis and sound segregation.* Ph.D., Groningen University, The Netherlands, 1983

[10] J.F. Culling and C.J. Darwin. Perceptual separation of simultaneous vowels: within and across-formant grouping by Fo. *Journal of the Acoustical Society of America, 93, pp 3454-3467,* 1993

[11] J.P.L. Brokx and S.G. Nooteboom. Intonation and the perceptual separation of simultaneous voices. *Journal of Phonetics, 10, pp 23-36,* 1982

[12] J. Bird and C.J. Darwin, Effects of a difference in fundamental frequency in separating two sentences in *Psychophysical and physiological advances in hearing.* A. R. Palmer, A. Rees, A. Q. Summerfield and R. Meddis(ed.), 1998, Whurr. London.

[13] S. McAdams. *Spectral fusion, spectral parsing and the formation of auditory images,.* unpublished Ph.D. dissertation, Stanford University, 1984

[14] C.J. Darwin and G.J. Sandell. Absence of effect of coherent frequency modulation on grouping a mistuned harmonic with a vowel. *Journal of the Acoustical Society of America, 97, pp 3135-3138,* 1995

[15] C.J. Darwin, V. Ciocca and G.R. Sandell. Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic. *Journal of the Acoustical Society of America, 95, pp 2631-2636,* 1994

[16] Q. Summerfield and J. Culling. Auditory segregation of competing voices: absence of effects of FM or AM coherence. *Philosophical Transactions of the Royal Society of London. Series B, 336, pp 357-366,* 1992

[17] R.P. Carlyon. Discriminating between coherent and incoherent frequency modulation of complex tones. *Journal of the Acoustical Society of America, 89, pp 329-340,* 1991

[18] M.H. Chalikia and A.S. Bregman. The perceptual segregation of simultaneous vowels with harmonic, shifted or random components. *Perception and Psychophysics, 53, pp 125-133,* 1993

[19] R.W. Hukin and C.J. Darwin. Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. *Perception and Psychophysics, 57, pp 191-196,* 1995

[20] R.M. Warren, C.J. Obusek and J.M. Ackroff. Auditory induction: perceptual synthesis of absent sounds. *Science, 176, pp 1149-1151,* 1972

[21] S. McAdams, M.C. Botte and C. Drake. Auditory continuity and loudness computation. *J Acoust Soc Am, 103, pp 1580-91,* 1998

[22] C.J. Darwin, Perceiving vowels in the presence of another sound: a quantitative test of the "Old-plus-New" heuristic. in *Levels in Speech Communication: Relations and Interactions : a tribute to Max Wajskop.* C. Sorin, J. Mariani, H. Méloni and J. Schoentgen(ed.), 1995, Elsevier. Amsterdam.

[23] M. Kubovy, Concurrent pitch segregation and the theory of indispensible attributes in *Perceptual Organization.* M. Kubovy and J. R. Pomerantz(ed.), 1981, Erlbaum. Hillsdale, N.J.

[24] M. Bodden. Auditory demonstrations of a cocktail-party processor. *Acustica, 82, pp 356-357,* 1996

[25] F.L. Wightman and D.J. Kistler. The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America, 91, pp 1648-1661,* 1992

[26] J.F. Culling and Q. Summerfield. Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. *Journal of the Acoustical Society of America, 98, pp 785-797,* 1995

[27] T.F. Munte, C. Kohlmetz, W. Nager and E. Altenmuller. Superior auditory spatial tuning in conductors. *Nature, 409, pp 580.,* 2001

[28] C.J. Spence and J. Driver. Covert spatial orienting in audition: exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 20, pp 555-574,* 1994

[29] W.A. Woods and S. Colburn. Test of a model of auditory object formation using intensity and interaural time difference discriminations. *Journal of the Acoustical Society of America, 91, pp 2894-2902,* 1992

[30] N.I. Hill and C.J. Darwin. Lateralisation of a perturbed harmonic: effects of onset asynchrony and mistuning. *Journal of the Acoustical Society of America, 100, pp 2352-2364,* 1996

[31] C.J. Darwin and R.W. Hukin. Auditory objects of attention: the role of interaural time-differences. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, *pp 617-629*, 1999

[32] C.J. Darwin and R.W. Hukin. Effectiveness of spatial cues, prosody and talker characteristics in selective attention. *Journal of the Acoustical Society of America*, *107*, *pp 970-977*, 2000

[33] C.J. Darwin and R.W. Hukin. Effects of reverberation on spatial, prosodic and vocal-tract size cues to selective attention. *Journal of the Acoustical Society of America*, *108*, *pp 335-342*, 2000

[34] R.P. Carlyon, R. Cusack, J.M. Foxton and R.H. Robertson. Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, *pp 115-127*, 2001