Paper for the 11th International Conference on Hearing, Grantham, UK, August, 1997

Effects of a difference in fundamental frequency in separating two sentences.

Jon Bird and C.J. Darwin
Experimental Psychology, University of Sussex
Falmer, Brighton, BN1 9QG, U.K.

1. Introduction

The perceptual separation of two competing voices is easier when the voices are on different fundamental frequencies (Fos). Brokx and Nooteboom (Brokx and Nooteboom, 1982; Brokx et al., 1979) asked their subjects to recall semantically anomalous sentences played against a continuous background of speech. They used LPC analysis and resynthesis to create a constant Fo difference between the monotone target sentence and the interfering speech. As the Fo difference between the two speech streams increased so did the intelligibility of the target sentence. The word recognition rate was about 40% for equal Fos and rose roughly linearly to about 60% for a difference of three semitones.

An increase in identification with a difference in Fo has also been found with a simpler experimental paradigm using pairs of steady-state synthesised vowels. Subjects identified both of a pair of simultaneous isolated vowels more accurately when they were on different Fos than when they were on the same Fo (Assmann and Summerfield, 1990; Culling and Darwin, 1993; Scheffers, 1983). Comparing the two sets of experiments it is clear that identification increases more rapidly with difference in Fo for synthetic vowels than for sentences. For vowels, most of the change occurs over the first semitone, with a substantial increase at only half a semitone.

Because of the simplicity of the paradigm, modelling and neuro-physiological investigations of speech separation by Fo differences have generally used "double-vowels" (Assmann and Summerfield, 1990; de Cheveigné, 1993; Meddis and Hewitt, 1992; Palmer, 1988; Summerfield and Culling, 1994). However, it is not clear whether the mechanisms responsible for the improvement in intelligibility at small Fo differences in double vowels are also generally useful for the separation of more natural speech.

Specifically, the improvement in identification of double vowels at very small (up to 1 semitone) Fo differences is likely to be due to the changes in amplitude at harmonic frequencies brought about by the beating of similar low-numbered harmonics in the F1 region of the two vowels (Assmann and Summerfield, 1994; Culling and Darwin, 1994). Although the fact that this mechanism can improve the intelligibility of double vowels is a tribute to the ingenuity of the auditory system in dealing with esoteric stimuli, the mechanism is unlikely to make a significant contribution to the separation of two natural voices where pitches and formant frequencies change continuously.

The perceptually more interesting mechanism whereby formants (or harmonics) from different frequency regions across the vowel spectrum are grouped together on the basis of a common pitch is evidently only used by listeners at larger Fo differences (Culling and Darwin, 1993) where performance on normal double vowel sounds has already asymptoted. The large Fo differences needed may be partly due to the relative insensitivity of the auditory system to pitch differences between sounds consisting of only unresolved harmonics (Carlyon, 1994; Houtsma and Smurzynski, 1990).

The two experiments reported here were modelled on Brokx and Nooteboom's (1982) experiments using resynthesised natural speech. They aimed to give more information about the mechanisms by which the auditory system exploits Fo differences in separating two sentence-length utterances.

Like Brokx and Nooteboom's experiments, they used resynthesised speech played on a monotone. However, they differed in using stimuli that were entirely voiced and had few stop consonants. By keeping the speech voiced and removing abrupt onsets and offsets, which could themselves act as grouping cues, we hoped to maximise the effect of Fo differences on intelligibility (as well as avoiding the problems of voicing errors in resynthesis).

2. Experiment 1

The first experiment, like Brokx and Nooteboom's (1982) used LPC resynthesised speech. It was designed to give an initial indication of whether the paradigm was feasible with entirely-voiced utterances, and specifically to test whether the change in intelligibility of the target sentence with a difference in Fo differed when the interfering sentence was higher or lower in pitch than the target.

2.1 Method

A set of 88 sentences which consisted solely of voiced phonemes (the consonants being mainly nasals and glides) read in a monotone voice (by JB) were stored digitally at 11.025kHz, after low-pass filtering at 5kHz. Each sentence was analysed and resynthesised by linear predictive coding (LPC). Each stimulus sentence pair consisted of a long and a short sentence added together with the short sentence roughly in the middle of the long sentence. All the short (target) sentences were resynthesised with an Fo of 140 Hz; the long (interfering) sentences were resynthesised at 0, ±1, ±2, ±4, ±6, or ±8 equal-tempered semitones from 140 Hz. Four different long sentences were resynthesised at each of these pitch differences.

The direction of the Fo difference was counter-balanced across two subject groups of 7 subjects each. Both groups heard the same stimuli with the same absolute Fo difference, but with a different polarity. For example, group A heard the long (interfering) sentence, "A normal animal will run away" at 132 Hz (one semitone below the target sentence) and subject group B heard the same sentence at 148 Hz (one semitone above the target sentence). Sentence pairs were presented in the same order to both subject groups. The experiment thus provides a counter-balanced comparison (within-subjects) between positive and negative Fo differences, but confounds the absolute size of the Fo difference with materials (4 sentence pairs per condition) and their order of presentation.

The subjects were instructed to listen to each stimulus twice in a particular trial. On the first hearing they were to listen to the long sentence and then write it down; on the second hearing they were to listen to the shorter sentence and then write it down. The subjects were not aware that only the short sentences would be used for analysis. The purpose of hearing the long sentence first was to give the subjects some idea of when the shorter sentence would begin, analogous to the warning in Brokx and Nooteboom's experiment. Subjects initially heard 5 practice stimuli to familiarise them with the experimental procedure. They could take as long as they wished to write their answers.

2.2 Results

The percentage (to control for variable numbers of words in a sentence) of words correctly reported was calculated for each sentence and each subject.

Recognition rate is a highly significant function of Fo difference ($F_{10,130} = 18.3$, $p < 0.0001$), but the change from zero is very similar for positive and negative Fo differences ($F_{4,52} = 0.4$, $p > 0.5$), consequently word recognition rate as a function of absolute Fo difference is shown in Figure 1, along with corresponding data from previous experiments. Our data show a much larger effect of Fo difference than do previous experiments. The particularly low performance with zero Fo difference is probably due to two factors: (i) the sentence task is harder with partial information than the double-vowel task, and (ii) continuously-voiced speech with no stop closures reduces the use of onset and offset cues to segregation.

A 1-semitone difference in Fo gave no significant increase in recognition (p=0.5), but a 2-semitone increase gave a significant increase over 1 of about 32% (p<0.0001). Recognition then stays constant until 6 semitones, before increasing again to 8 semitones (p=0.0003). The large increase in recognition with Fo differences thus occurs later in the present experiment than in double vowel experiments. There is no significant increase in recognition at 1 semitone in the present experiment; in contrast, double vowel experiments show no increase in recognition after 1 semitone. The present experiment also shows a continuing increase in recognition from 6 to 8 semitones.

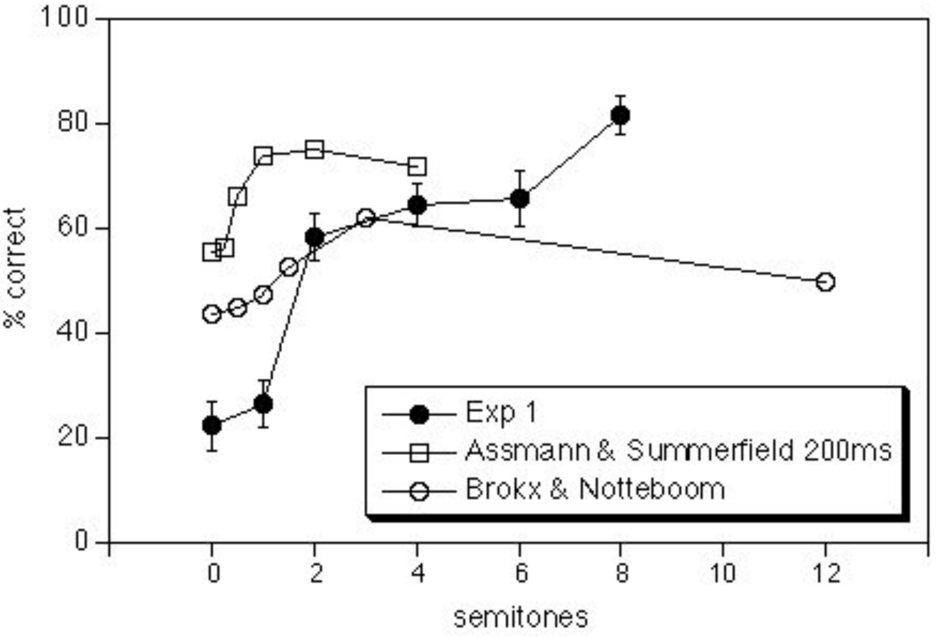

Figure 1. Results from experiment 1 as a function of absolute Fo difference compared with results from previous experiments on speech separation. Error bars are ±1 standard error of the mean.

3. Experiment 2

The second experiment investigated the mechanisms responsible for the increase in intelligibility with Fo found in experiment 1. In particular it used similar stimulus manipulations to those introduced by Culling and Darwin (1993) to separate mechanisms that use a difference in Fo: (i) within a frequency region, e.g. to improve formant frequency estimation; (ii) to group components globally across different frequency regions.

Culling & Darwin split the spectrum of individual vowels into low-frequency and high-frequency parts mid-way between their F1 and F2 frequencies and synthesised each part either on the same or on different Fos. Pairs of vowels could then be combined so that, for example, the low-frequency part of one vowel had the same Fo as the high-frequency part of another vowel (Fo-swapped). They showed that, for differences in Fo up to a semitone, only the low-frequency region contributed to the improvement in double-vowel recognition. Much of this improvement could be attributed to dynamic amplitude changes in individual harmonics as a result of beating (Assmann and Summerfield, 1994; Culling and Summerfield, 1995). Only with larger differences in Fo, was vowel identification influenced by Fo differences in the high-frequency region.

This experiment counterbalanced the size of the Fo difference with materials in order to test the indication from Experiment 1 that most of the improvement in recognition for sentence-length speech occurred after 1 semitone. It also asked whether this improvement was due to global grouping of components on the same Fo are to mechanisms restricted to particular frequency regions.

3.1 Method

The stimulus preparation for Experiment 2 differed from that of Experiment 1 in the following ways. 80 of the previous 88 sentences were read by a different speaker (CJD). PSOLA was used rather than LPC to produce higher-quality resynthesis on five different Fos : 100 Hz, 105.94 Hz, 112.25 Hz, 133.49 Hz and 178.18 Hz (0, 1, 2, 5 and 10 semitones above 100 Hz). Each resynthesised sentence was filtered (501-point FIR) at 800 to produce both a low- and high-pass part. A filter cut-off of 800 Hz was chosen to ensure resolved harmonics in the low-pass region and to give roughly equal intelligibility for the low- and high-pass parts.

Sentences were paired in four conditions, each of which had five Fo differences. Each pair consisted of a short target sentence and a longer interfering sentence which began 400 ms before and finished between 10 and 580 ms after the shorter. The conditions were:

i. **normal**- the target and interfering sentences each had a consistent Fo. The shorter, target sentence always had an Fo of 100 Hz, the longer, interfering sentence had a variable Fo.

ii. **Fo-swapped**- the low-pass part of the target sentence had the same Fo, 100 Hz, as the high-pass part of the interfering sentence, the other two parts shared the same, variable Fo. Mechanisms local to a particular frequency region should be unaffected by this manipulation, but global mechanisms should be impaired since they would produce inappropriate pairings of low- and high-pass parts.

iii. **same Fo in low-pass** - the low-pass part of both sentences shared the same Fo of 100 Hz, which was also shared with the high-pass part of the target sentence. The high-pass part of the interfering sentence had a variable Fo. This condition estimated the segregation arising from an Fo difference in the high-frequency region.

iv. **same Fo in high-pass** - the high-pass part of both sentences shared the same Fo of 100 Hz, which was also shared with the low-pass part of the target sentence. The low-pass part of the interfering sentence had a variable Fo. This condition estimated the segregation arising from an Fo difference in the low-frequency region.

The 40 sentence pairs were presented in the same order to 40 subjects, but different subjects heard a particular sentence pair in a different one of the 20 conditions. Rotating conditions across sentence pairs controlled for the confounding of particular combinations of conditions and sentence pairs that was present in Experiment 1.

Subjects listened to each stimulus once only, and were instructed to attend to and write down the sentence that started second.

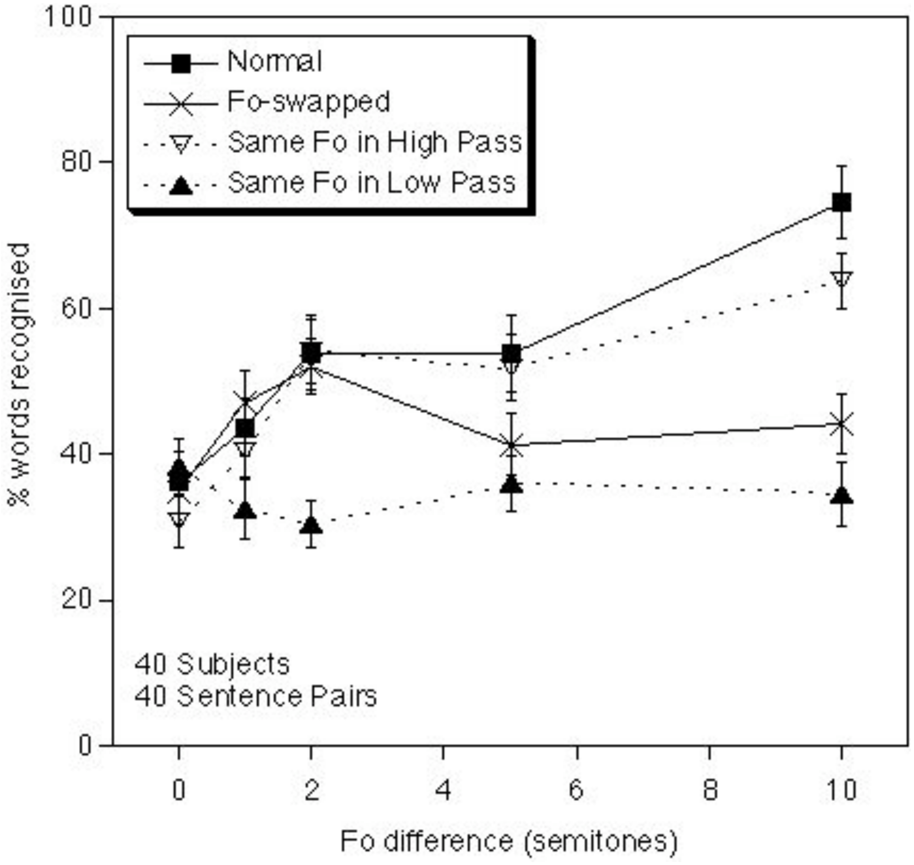

Figure 2. Percent of words recognised correctly as a function of Fo difference in the four conditions of

experiment 2. Error bars are ±1 standard error of the mean.

3.2 Results

The mean recognition rates are shown in Figure 2. Overall they show the main features of Culling & Darwin's (1993) results with double-vowel stimuli: namely, that the improvement for small Fo differences is due almost exclusively to Fo differences in the low frequency region. They differ from the double-vowel results in that recognition continues to improve beyond one semitone - the Fo difference at which double-vowel identification asymptotes.

Looking at the results in closer detail, there are clear differences across conditions in the effect of an Fo difference ($F_{12,468}=4.5$, $p<0.0001$). The **Normal** condition shows no significant improvement between zero and 1 semitone ($p>0.2$), but a significant improvement between zero and 2 semitones ($p<0.005$). Recognition also improves significantly from 5 to 10 semitones ($p<0.001$). By contrast, although the **Fo-swapped** condition shows improvement over 1 ($p<0.05$) and 2 ($p<.005$) semitones, recognition then decreases so that the 5- and 10-semitone conditions do not differ from the zero-semitone condition. Since grouping by Fo across the low and high frequency regions is beneficial in the Normal condition, but detrimental in the Fo-swapped condition, the lack of difference between these two conditions at 0, 1 and 2 semitones and the significant difference at 5 ($p<0.05$) and 10 ($p<0.0001$) semitones indicates that such across-frequency grouping is only important for Fo differences greater than 2 semitones.

The substantial improvement up to 2 semitones is due almost entirely to factors local to the low-frequency region. These factors could include improved accuracy at identifying separate formants in the first-formant region for the two sentences, and correctly attributing then, on the basis of Fo, to the target or the interfering sentence. The **same Fo in high-pass** condition, which is physically identical to the **Normal** condition in the low-pass region gives very similar recognition to the **Normal** condition except at 10 semitones where it is marginally ($p<0.1$) worse. By contrast, the **same Fo in low-pass** condition, shows no significant improvement at all with Fo differences ($F<1$).

It is interesting that recognition in the the **Fo-swapped** condition is worse than in the **same Fo in high pass** condition at 5 and 10 semitones. It is apparently harder to incorporate the correct high-pass region on the wrong Fo than it is to reject the wrong high-pass region on the correct Fo.

In order to get a more stable estimate of the effects of small Fo differences on recognition, data from the **Normal**, **Fo-swapped** and **same Fo in high-pass** conditions were analysed for differences of 0, 1 and 2 semitones. The three conditions did not differ ($F<1$), and together gave a linear increase in recognition rates from 0 (34%) to 1 (44%) to 2 (53%) semitones with both steps significant at $p<0.01$. The lack of improvement in Experiment 1 from zero to 1 semitone is therefore not replicated and could have been due to confounding sentence pairs with particular Fo differences.

As discussed in the Introduction, with double-vowel stimuli the improvement in identification for small Fo differences, is probably substantially due to spectral changes induced by beating between corresponding harmonics from the two vowels. It is not obvious that a similar explanation is appropriate for the present results. With sentence material it is more likely that at any one time a harmonic from one voice will dominate the neighbouring harmonic from the other. Consequently, it may be possible, at 1 semitone difference, to resolve an individual component and attribute it to either the target or masking sentence. With a 2-semitone difference (12%) two neighbouring components may be resolvable even when they are of similar amplitude. Further experiments are needed to clarify the mechanism involved.

Two differnces between the results of Experiments 1 and 2 are probably due to changes in experimental design. The better recognition in the present experiment than in Experiment 1 when there was no Fo-difference may be due to the higher quality of the PSOLA than the LPC resynthesis and the change in speaker. The smaller overall improvement with Fo in Experiment 2 could be due to listeners only hearing the sentences once on a particular trial rather than twice in Experiment 1.

4. Summary

In summary, the two experiments reported here have:

- confirmed Brokx & Nooteboom's (1982) findings that a difference in Fo continues to improve the recognition of one sentence masked by another beyond the 1-semitone difference at which double-vowel identification asymptotes;
- extended to sentence recognition and out to 2 semitones the finding by Culling & Darwin (1993) that such improvement arises from mechanisms confined to the low-frequency (<800 Hz) region;
- shown that use of a common Fo to group components across the low- and high-frequency regions occurs for differences in Fo of 5 semitones and above, but not for 2 semitones and below.

Acknowledgements

Jon Bird carried out the first experiment while holding a Nuffield Foundation Undergraduate Research Bursary and the second while holding a BBSRC research studentship. We would like to thank Paul Russell for porting the PSOLA software to MPW, Rob Hukin for technical assistance, Gustav Kuhn for help in running Experiment 2, and the MRC for grant G9505738N.

References

Assmann, P.F. and Summerfield, A.Q. (1990) Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. J. Acoust. Soc. Am. 88, 680-697.

Assmann, P.F. and Summerfield, A.Q. (1994) The contribution of waveform interactions to the perception of concurrent vowels. J. Acoust. Soc. Am. 95, 471-484.

Brokx, J.P.L. and Nooteboom, S.G. (1982) Intonation and the perceptual separation of simultaneous voices. J. Phon. 10, 23-36.

Brokx, J.P.L., Nooteboom, S.G. and Cohen, A. (1979) Pitch differences and the integrity of speech masked by speech. IPO Annual Progress Report 14, 55-60.

Carlyon, R.P. (1994) Detecting pitch-pulse asynchronies and differences in fundamental frequency. J. Acoust. Soc. Am. 95, 962-965.

Culling, J.F. and Darwin, C.J. (1993) Perceptual separation of simultaneous vowels: within and across-formant grouping by Fo. J. Acoust. Soc. Am. 93, 3454-3467.

Culling, J.F. and Darwin, C.J. (1994) Perceptual and computational separation of simultaneous vowels: cues arising from low frequency beating. J. Acoust. Soc. Am. 95, 1559 - 1569.

Culling, J.F. and Summerfield, Q. (1995) Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. J. Acoust. Soc. Am. 98, 785-797.

de Cheveigné, A. (1993) Separation of concurrent harmonic sounds: fundamental frequency estimation and a cancellation model of auditory processing. J. Acoust. Soc. Am. 93, 3271-3290.

Houtsma, A.J.M. and Smurzynski, J. (1990) Pitch identification and discrimination for complex tones with many harmonics. Journal of the Acoustical Society of America 87, 304-310.

Meddis, R. and Hewitt, M. (1992) Modelling the identification of concurrent vowels with different fundamental frequencies,. J. Acoust. Soc. Am. 91, 233-245.

Palmer, A.R. (1988) The representation of concurrent vowels in the temporal discharge patterns of auditory nerve fibers. In H. Duifhuis, J. W. Jorst and H. P. Wit (Ed.), Basic Issue in Hearing, pp. 244-251, Academic, London.

Summerfield, Q. and Culling, J.F. (1994) Auditory computations that separate speech from competing sounds: a comparison of monaural and binaural processes. In E. Keller (Ed.), Fundamentals of speech synthesis and speech recognition, pp. 313-338, Wiley, Chichester.

20 March 1997