

FUNDAMENTAL FREQUENCY AND THE INTELLIGIBILITY OF COMPETING VOICES

Peter F. Assmann

School of Human Development, The University of Texas at Dallas, Richardson TX 75083

ABSTRACT

When two people speak at the same time, it is easier to understand what either of them is saying if the voices differ in fundamental frequency (F_0). The contribution of F_0 to the intelligibility of pairs of simultaneous sentences was investigated using a high-quality speech vocoder. Word identification accuracy improved by 23% as the F_0 difference between the voices increased from 0 to 8 semitones. With a 4-semitone difference, benefits of F_0 were asymmetrical: the sentence with the higher F_0 was more intelligible. F_0 differences provided similar gains when F_0 was held constant using a monotone pitch, and when the natural pattern of F_0 variation was preserved in each sentence. Time-varying F_0 did not result in improved identification compared to constant F_0 , except at 0 and 1 semitones where marginal improvements could be attributed to perceptual segregation based on momentary differences in static F_0 , rather than F_0 dynamics.

1. INTRODUCTION

Listeners rely on several perceptual strategies to separate a target voice from competing sounds. When the competition is created by a second voice, listeners find it easier to understand the message if the competing voice has a different fundamental frequency (F_0). The benefits of F_0 differences (ΔF_0 's) for sentence intelligibility were demonstrated experimentally by Brox and Nootboom [1]. They recorded 96 semantically anomalous Dutch sentences (e.g. "*The town swims now in a sheep*") spoken by an adult male, along with a 600-word short story which served as the masker. The masker was produced as a continuous stream of speech, and was edited to remove silent intervals. Linear predictive coding (LPC) was then used to resynthesize both the target sentences and masker passage on a constant (monotone) pitch. The F_0 of the masker was fixed at 100 Hz, while the target sentences were between 0 and 12 semitones higher. Brox and Nootboom showed that sentence intelligibility (percentage of target words correctly identified) was higher when the target and masker sentences had different F_0 's. Performance improved as the ΔF_0 increased from 0 to 3 semitones, but dropped with a ΔF_0 of 12 semitones (one octave). They attributed the benefits of ΔF_0 's to perceptual segregation based on the difference in pitch, combined with a release from perceptual fusion: the tendency for two simultaneous sounds to blend into one when they have identical pitches [2].

Bird and Darwin [3] extended these results in two experiments with pairs of sentences made up entirely of voiced sounds (e.g. "*A normal animal will run away*"). In each pair they embedded a short (target) sentence within a longer (masker) sentence. In their first experiment, which used LPC analysis-synthesis, they found dramatic improvement in the identification

of the target sentences as the ΔF_0 increased from 0 to ± 8 semitones, relative to a baseline F_0 of 140 Hz. In a second experiment they used a PSOLA speech coder to determine the relative contributions of the low-frequency (0-800 Hz) and high-frequency (>800 Hz) portions of the speech spectrum to the ΔF_0 effect. They found very little change when the ΔF_0 was eliminated in the high-frequency region but retained at low frequencies. However, when ΔF_0 's were eliminated from the low-frequency region the benefits for identification disappeared. They found a mixed pattern when they "swapped" the F_0 's across the low- and high-frequency portions of the two sentences. For small ΔF_0 's (up to 2 semitones) the F_0 -swapped condition gave the same pattern of improvement as the baseline condition, despite the mismatch that would result from grouping together low-frequency and high-frequency parts of the spectrum on the basis of common F_0 . However, with larger ΔF_0 's (5 and 10 semitones) the F_0 -swapped condition gave no improvement, while the baseline condition increased by as much as 35%. Bird and Darwin attributed the benefits with large ΔF_0 's to a process of across-frequency grouping based on common F_0 . They suggested that other mechanisms that depend mainly on cues in the low-frequency region may contribute with smaller ΔF_0 's.

When two voices compete, listeners with normal hearing can exploit temporal fluctuations in overall amplitude in a masker sentence to "glimpse" words and fragments of the target sentence [4,5]. These opportunities for glimpsing are reduced when the maskers are edited to remove silent intervals as in [1] or when sentences are constrained to have only voiced sounds as in [3]. The restriction to use exclusively voiced segments was intended to minimize the contribution of additional grouping cues, such as those provided by abrupt changes in voicing and sudden onsets and offsets. However, by restricting opportunities for glimpsing, these experiments may have overestimated the contribution of ΔF_0 's in connected speech, in which a substantial portion of the waveform is composed of voiceless sounds for which ΔF_0 's are not available. The sentences used in the experiment described below contained both voiced and voiceless segments, without editing of maskers.

Both Brox and Nootboom [1] and Bird and Darwin [3] synthesized the sentences with a constant F_0 , producing a monotone voice pitch and eliminating natural time-varying changes in F_0 . There are at least four reasons why F_0 variation over time might aid intelligibility: (i) momentary differences in F_0 promote source segregation [6]; (ii) coherent F_0 modulation provides a basis for tracking a voice over time; (iii) variation in F_0 over time reduces overlap and perceptual fusion; (iv) linguistic knowledge of the direction and extent of pitch changes (prosody) can help to disambiguate words in sentences that are masked or

distorted by the sounds of a competing voice.

The present study extended Bird and Darwin's paradigm to determine the separate contributions of ΔF_0 's and F_0 variation across time. Sentences were synthesized either with a constant F_0 (monotone sentences) or with natural, time-varying changes in F_0 (intoned sentences) in which the trajectory of measured F_0 's was shifted to have the same average as the corresponding monotone sentence. Sentences in the monotone and intoned conditions had the same average baseline F_0 of 100 Hz, but F_0 variation over time in the latter condition typically ranged between about 80 and 120 Hz (± 3 -4 semitones) around the 100 Hz baseline.

2. METHOD

A set of 48 "high-predictability" sentences from the SPIN (Speech Perception in Noise) test [7] was recorded by an adult male speaker from north central Texas. The stimuli were recorded on DAT tape, transferred to disk via a digital interface at a 48 kHz sample rate, downsampled to 24 kHz and processed using the STRAIGHT speech analysis-synthesis program [8,9]. F_0 was estimated at 1-ms intervals, and each frame was identified as voiced or unvoiced. In monotone sentences, the measured F_0 in each voiced frame was replaced by a constant F_0 of 100 Hz, or was 1, 2, 4, 6, or 8 semitones higher (106, 112, 126, 143, or 160 Hz). In intoned sentences the natural pattern of variation in F_0 was preserved by shifting the measured F_0 pattern along the frequency scale to generate the same average F_0 difference. Measured F_0 values were shifted by a constant amount to give the same mean F_0 (averaging across all voiced frames) as the corresponding monotone sentence. Sentences were combined in pairs at a nominal 0 dB target-to-masker ratio and were temporally aligned at their offsets. In each pair, one member had a mean F_0 of 100 Hz, while the other was 0-8 semitones higher. An example is shown in Fig 1.

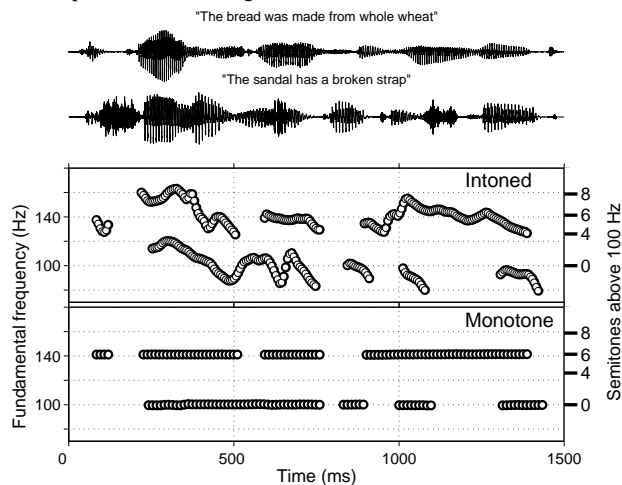


Figure 1: Example of a sentence pair with a ΔF_0 of 6 semitones (100 & 141 Hz). The waveforms of the two original sentences are shown at the top. The lower panels show the F_0 trajectories used to synthesize the monotone and intoned versions of the two sentences.

Listeners were undergraduate Psychology students with little or no previous experience listening to synthetic speech. All were

native speakers of American English and reported normal hearing. They were tested individually in a double-walled sound booth. Sentence pairs were presented monaurally over headphones at a mean level of 80 dB SPL (A) and responses were typed at a computer keyboard. Each sentence pair was presented twice. Listeners were instructed to type as many words as possible from either sentence; on the second presentation, they were asked to type the remaining words, including any missed the first time, in any order.

Two separate groups of 18 subjects were randomly assigned to two different order conditions: half completed the monotone sentences first, while the other half began with the intoned sentences. Prior to the main experiment each subject completed a practice session (with feedback) in which they heard 10 pairs of sentences similar to the set used in the main experiment, but spoken by a different male talker. Following the practice set they completed 2 sets of 24 trials in the main experiment without feedback.

3. RESULTS

Identification accuracy was measured in terms of the number of keywords in each sentence identified correctly. The number of keywords per sentence ranged from 3 to 6 and hence the scores were expressed as percentages.

3.1 Effects of ΔF_0 and F_0 variation

An analysis of variance indicated a significant main effect of ΔF_0 [$F(5,85)=14.57$; $p<0.01$] but the effect of F_0 variation was not significant [$F(1,17)=1.99$; $p=0.18$], nor was there an interaction of these two variables [$F(5,85)=0.65$; $p=0.66$]. Figure 2 shows that identification accuracy improved as a function of ΔF_0 , with a similar pattern of improvement for constant and variable F_0 's.

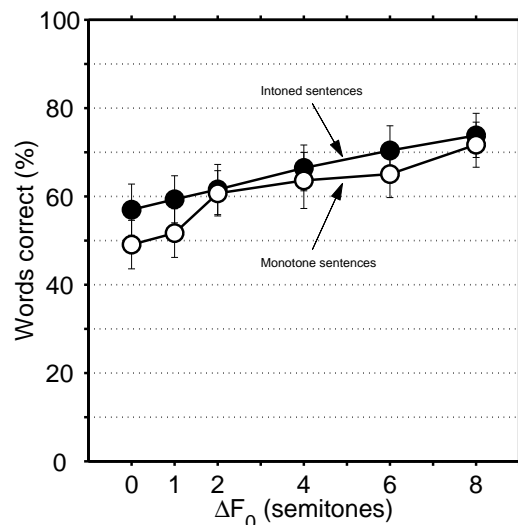


Figure 2: Sentence intelligibility (percent words correct) as a function of ΔF_0 . Unfilled circles show the results for monotone sentences (constant F_0 condition); filled circles show the results for the intoned sentences (variable F_0 condition).

Consistent with [1] and [3] listeners showed a continuous

improvement in performance as a function of ΔF_0 . Baseline performance (at a ΔF_0 of 0 semitones) was higher than that obtained in Bird and Darwin's first experiment, but the maximum benefit of ΔF_0 (comparing 0 and 8 semitones) was smaller (23% compared to >50%). The higher baseline performance and reduced benefit of ΔF_0 may stem from differences in the synthesis method and the stimulus materials, particularly the inclusion of voiceless segments which may engage other forms of perceptual grouping and glimpsing processes.

Although not statistically significant, mean scores were about 8% higher for intoned sentences than monotone sentences at 0 and 1 semitones. Performance is expected to be better with a ΔF_0 of 0 semitones for intoned sentences because F_0 variation over the time course of the sentence leads to momentary differences in F_0 . The standard deviation in measured F_0 over the course of a sentence was about 12 Hz (or 2 semitones, relative to the mean F_0). Hence it seems likely that this factor, rather than listeners' sensitivity to F_0 modulation, contributed to the small increase in identification accuracy for intoned sentences in this experiment.

3.2 Lower versus higher F_0

The main effect of relative F_0 (lower versus higher) was not significant, but there was an interaction of relative F_0 and ΔF_0 [$F_{(5,85)}=3.00$; $p<0.05$]. Post-hoc tests revealed more accurate identification of the sentence with the higher F_0 than the lower F_0 when the ΔF_0 was 4 semitones. The interaction is shown in Figure 3.

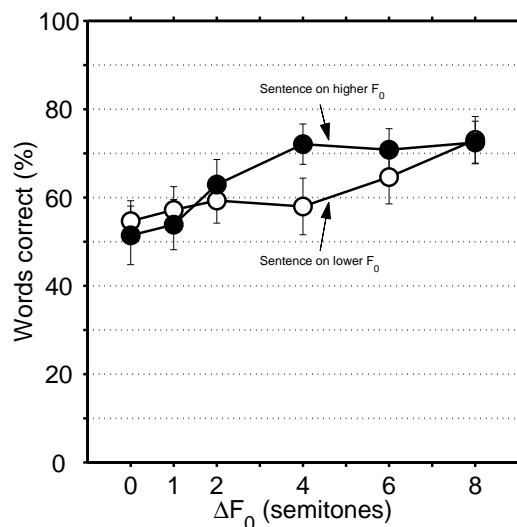


Figure 3: Sentence intelligibility (percent words correct) as a function of ΔF_0 . Unfilled circles show the results for the sentence in each pair with the lower F_0 ; filled circles show the results for the sentence with the higher F_0 .

The increased intelligibility of the sentence with the higher F_0 may be related to differences in the relative perceptual salience of the pitches of pairs of simultaneous vowels. When two vowels are presented concurrently on different F_0 's, listeners identify

them more accurately if they differ in F_0 [10] and they can often report which vowel has the higher pitch [11]. When asked to match the pitches of the two vowels, they frequently judge the vowel with the higher F_0 to have the dominant pitch, and tend to match its F_0 more consistently and accurately than the lower F_0 . The tendency for the higher-pitched vowel to dominate was strongest with a ΔF_0 of 4 semitones, the largest ΔF_0 included in the study [12]. The increased intelligibility of the higher-pitched sentence at a ΔF_0 of 4 semitones may be linked to the enhanced salience of its pitch. Further work is needed to determine the perceptual basis for the asymmetry.

4. DISCUSSION

The results confirm that ΔF_0 is a powerful cue for the perceptual segregation of competing voices. The benefits of F_0 were smaller in the present study (a 23% improvement from 0 to 8 semitones, compared to >50% in Bird and Darwin's first experiment). Bird and Darwin's sentences contained only voiced sounds, while the sentences used here included both voiced and voiceless sounds, thereby increasing the opportunities for glimpsing which are more representative of natural speech.

The absence of an effect of F_0 variation is surprising, given the diverse reasons why it might be expected to lead to improved identification. Nonetheless, the results are consistent with studies of the perception of concurrent vowels [13] which suggest that F_0 modulation can increase the perceptual salience of a vowel against the background of another vowel, but does not provide a basis for improved vowel identification via mechanisms of across-frequency grouping and segregation. On the other hand, unlike the sinusoidal variation used to study the role of frequency modulation in concurrent vowels, F_0 variation in natural speech is complex, and has the potential to contribute to the perceptual segregation of voices in a variety of ways. Further investigations with multiple voices, a range of target-to-masker amplitude ratios, and additional prosodic contexts may reveal them.

5. CONCLUSIONS

Consistent with [1,3] listeners identified words in pairs of simultaneous sentences more accurately when they were synthesized with different F_0 's. Identification accuracy improved by 23% when the F_0 difference was increased from 0 to 8 semitones (100/160 Hz). When the voices differed by 4 semitones (100/126 Hz) the voice with the higher F_0 was more intelligible. No evidence was found that listeners can exploit the natural time-varying changes in F_0 to track a voice through the background of a second voice speaking at a similar level.

ACKNOWLEDGMENTS

I am grateful to Hideki Kawahara for the providing the software to implement the STRAIGHT synthesizer, to Dwayne Paschall for helpful discussions, and to Sneha Bharadwaj for assistance in conducting the listening tests.

REFERENCES

- [1] Brox, J. P. L. and Nootboom, S. G. 1982. Intonation and the perception of simultaneous voices. *Journal of Phonetics*, 10, 23-26.
- [2] Stumpf, C. 1890. *Tonpsychologie*. Leipzig: S. Hirzel-Verlag. (English translation of §19, Bd. II by B. Rand, In B. Rand (ed.), *The Classical*

Psychologists; Selections Illustrating Psychology from Anaxagoras to Wundt. Magnolia : P. Smith Publ., Inc., 1912).

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

[4] Miller, G.A. 1947. The masking of speech. *Psychol. Bulletin*, 44, 105-129.

[5] Festen, J.M. and Plomp, R. 1990. Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.*, 88, 1725-1736.

[6] Bregman, A.S. 1990. *Auditory scene analysis*. Cambridge, MIT Press.

[7] Kalikow, D. N., Stevens, K. N., and Elliott, L. L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.* 61, 1337-1351.

[8] Kawahara, H. 1997a. STRAIGHT-TEMPO: a universal tool to manipulate linguistic para-linguistic speech information. *IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, Vol. 2, pp. 1620-1625.

[9] Kawahara, H. 1997b. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 1303-1306.

[10] Scheffers, M. T. 1983. *Sifting vowels: auditory pitch analysis and sound segregation*. Doctoral dissertation, University of Groningen, The Netherlands.

[11] Paschall, D.D. and Assmann, P.F. 1998. Ranking the pitches of concurrent vowels. *Proceedings of the 16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America*, Vol. 3, pp. 2009-2010.

[12] Assmann, P. F. and Paschall, D. D. 1998. Pitches of concurrent vowels. *J. Acoust. Soc. Am.* 103, 1150-1160.

[13] Culling, J. F. and Summerfield, Q. 1995. The role of frequency modulation in the perceptual segregation of competing vowels. *J. Acoust. Soc. Am.* 98, 837-846.