# Effect of prosodic changes on speech intelligibility

*Catherine Mayo*[1], *Vincent Aubanel*[2,3], *Martin Cooke*[2,3]

[1]Centre for Speech Technology Research, University of Edinburgh, Edinburgh UK
[2]Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
[3]Ikerbasque (Basque Science Foundation), Bilbao, Spain

catherin@inf.ed.ac.uk, v.aubanel@laslab.org, m.cooke@ikerbasque.org

## Abstract

Talkers adopt different speech styles in response to factors such as the perceived needs of the interlocutor, environmental noise and explicit instruction. Some styles have been shown to be beneficial for listeners but many aspects of the relationship between speech modifications and intelligibility remain unclear, particularly for prosodic changes. The current study measures the relative intelligibility in noise of speech spoken in 5 speech styles – plain, infant-, computer- and foreigner-directed, and shouted – and relates listener scores to acoustic/prosodic parameters and quantitative estimates of energetic masking. Intelligibility changes over plain speech correlated well with durational modifications, which included elongations of all segments as well as increases in the number of unfilled pauses. Both mean fundamental frequency and its range displayed great variation across styles but with no clear intelligibility benefits. Energetic masking per unit time was similar in each style but the total amount of speech which escaped masking was a good predictor of word identification rate. These findings suggest that much of the prosody-related intelligibility gain is derived from durational increases.

**Index Terms**: speech styles, prosody, intelligibility

## 1. Introduction

Clear speech—that is, speech produced for listeners with perceptual and/or linguistic deficits (e.g., hearing-impaired listeners, listeners in noisy environments, second language learners)—is characterised in the prosodic domain by changes to both speech rate and fundamental frequency ($F0$) compared to conversational or 'plain' speech (see review in [1]). Specifically, clear speech has been found to have a slower speech rate than plain speech, [2, 3] due to an increased number and duration of pauses [4, 5, 6, 3], and a non-linear increase in the duration of many segments [7, 3]. Clear speech also shows increased average $F0$ and increased $F0$ range [4].

Accounts such as the Hyper-Hypo (H&H) theory of speech production [8] suggest that these prosodic changes, along with other articulatory-acoustic changes associated with clear speech, are made by the talker to improve speech intelligibility for the listener (while maintaining minimum talker effort). Indeed, clear speech has been found to be more intelligible than plain speech when both speech types are presented in challenging listening conditions, e.g., [9]. However, it is not clear to what extent the *prosodic* changes seen in clear speech are actually attended to by listeners. Studies that have specifically examined the role of speech rate and $F0$ in improved speech intelligibility have found mixed results [10, 11, 12]. For example, some studies examining inter-talker variability found correlations between speech rate and talker intelligibility [7, 10], while others have found no such correlation [13, 14]. Similarly, studies in which speech was artificially scaled to make overall speech rate higher or lower had contradictory results with respect to changes in intelligibility [15, 16, 9]. Additionally, although it has been demonstrated that talkers can be trained to produce clear speech at plain speech rates [17], this quicker clear speech showed only a small, non-significant intelligibility advantage over plain speech for hearing-impaired listeners, suggesting that "for some listeners speaking rate is a contributing factor to the high intelligibility of clear speech" [1, p.221]. At the level of pauses, correlations have been demonstrated between number of pauses and speech intelligibility [3, 4]. However, the artificial insertion of pauses into plain speech was not found to increase intelligibility [9]. Finally, although artificially flattening $F0$ range has been found to decrease intelligibility [12, 18], artificially increasing average $F0$ [11] or average $F0$ and $F0$ range [19] was found not to increase intelligibility. Furthermore, in studies of inter-talker variability, average $F0$ was not correlated with talker intelligibility [7, 14, 10].

The question remains, therefore: what is the relationship, if any, between the prosodic adjustments found in clear speech as compared to plain speech, and the increased intelligibility found for such speech? In the current study we examine multiple characteristics of speech rate and $F0$, and relate these to subjective and objective intelligibility-in-noise measurements, with the aim of more fully describing this relationship.

## 2. Method

### 2.1. Stimuli

A male, native British English talker was recorded producing 25 randomly selected TIMIT sentences in five speech styles: plain, infant-, computer- and foreigner-directed, and shouted. All but the shouted speech was elicited via instructions designed to induce different speech styles in the talker (e.g., "Speak as if you were talking to a computer"). The resulting speech should therefore not be taken as representative of speech directed at a real interlocutor. The shouted speech was produced while the talker listened to 9-speaker babble-shaped noise [20] presented at an intense level.

Figure 1 shows spectrograms of one sentence produced in each of the five speech styles, showing clear durational differences between styles as well as in pitch patterns.

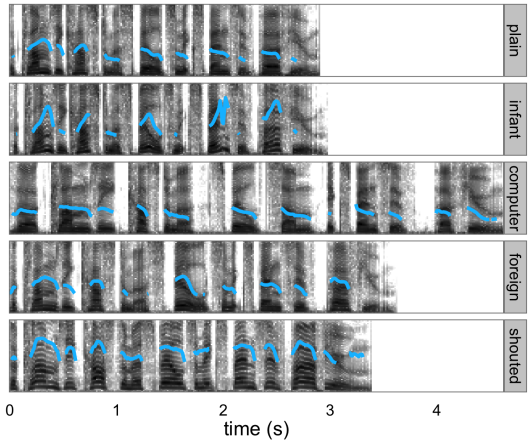


Figure 1: *Spectrograms for the sentence "The clumsy customer spilled some expensive perfume". Frequency range is 0–4 kHz, and overlaid $F0$ range is 88–310 Hz.*

All recorded sentences (25 sentences x 5 speech styles) were mixed with 9-speaker babble-shaped noise [20] at a signal-to-noise ratio (SNR) of 0 dB.

### 2.2. Subjective measurements

Twenty laboratory-based listeners and 60 listeners participating via the Amazon Mechanical Turk (AMT) crowdsourcing system performed intelligibility tests on the sentence-plus-noise mixtures. The laboratory-based listeners were all native British English talkers with no reported history of speech/language disorders. This group listened to the stimuli in sound-treated listening booths through headphones. The AMT listeners were also self-reported native English talkers (no accent specified) with no reported speech/language disorders. This group was instructed to listen to the stimuli over headphones in a quiet listening environment; these parameters could, however, not be controlled.

### 2.3. Objective measurements

Sentences were manually segmented at the phoneme level by a trained phonetician. Segments were subsequently grouped into nine sound classes, including unfilled pauses (see labels in Figure 4). $F0$ was extracted every 10 ms with PRAAT after adjusting the detection range to the speaker. Median and range was calculated for the voiced part of the signal, and range was taken as the interquartile range.

A glimpsing analysis [21] was performed to measure both the proportion and absolute number of time-frequency regions where the speech signal was sufficiently more energetic than the masker. Modelled spectro-temporal excitation patterns were computed for both speech and noise by processing the signal through a bank of 58 gammatone filters with centre frequencies in the range 100-8000 Hz, followed by log-compression of the smoothed Hilbert envelope in each band. The *glimpse count* is the number of time-frequency regions where the speech exceeds the noise by 3 dB, while *glimpse proportion* is a durational-normalised form.

## 3. Results

Results from both sets of listeners indicate that this talker produced computer-directed speech that was significantly more intelligible than all other types. Figure 2 shows the Word Error Rate for both sets of listeners across the five speech types. Correlation between the two groups is $[\rho = 0.95, p < .05]$. All listeners found infant-directed speech to be less intelligible than all other speech types; for AMT listeners (who showed significantly higher word error rates than lab-based listeners) this difference in intelligibility was significant.

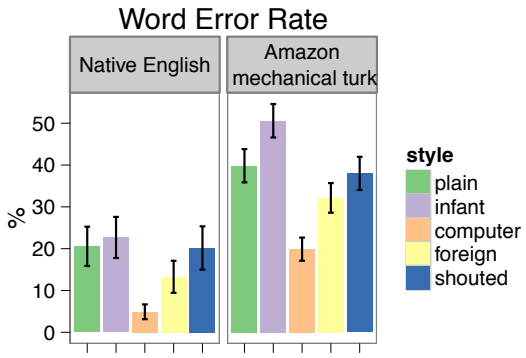


Figure 2: *Word Error Rate for Native English and Amazon Mechanical Turk listeners across the five speech types. Error bars, here and elsewhere, show 95% confidence intervals.*

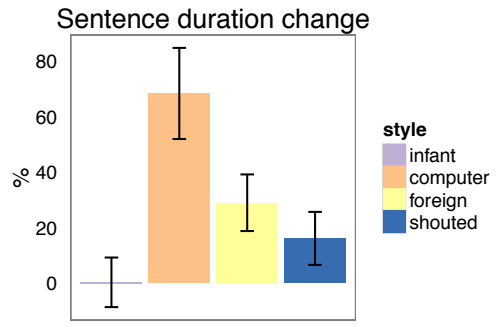


Figure 3: *Average sentence duration change relative to plain speech across sentences. Baseline is 2.87 seconds.*

Acoustic analysis of the recorded speech provides some possible explanations for these results. Sentence duration was significantly different across speech types [$p<.001$] All non-plain speech styles displayed longer overall sentence durations than plain speech, with the exception of infant-directed speech (Figure 3). Computer-directed sentences were 68.4% longer on average than plain speech sentences. Correlation with sentence duration and WER averaged over the two groups of listeners is [$\rho = -0.95, p < .05$].

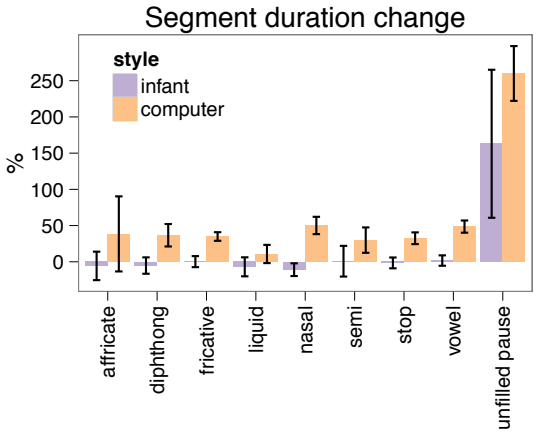


Figure 4: *Average segment duration change over segment classes across sentences.*

This increase in overall sentence duration is reflected in individual segment durations: separate ANOVAs for each segment class showed that segment durations varied significantly across speech types for *diphthong*, *fricative*, *nasal*, *stop*, and *vowel* [all $p<.001$] but were identical for *affricate*, *liquid* and *semi*. Figure 4 shows the duration changes relative to plain speech for infant- and computer-directed speech. Computer-directed speech segment duration increases were significant for all factors cited above, with greatest duration increases for nasals (50%) and vowels (49%). In contrast, infant-directed speech segment duration were statistically identical for all segment types. Unfilled pauses also varied across speech types [$p<.001$], being significantly longer in all non-plain speech types. Unfilled pause duration does not correlate well with WER across speech styles [$\rho = -0.55, p = .33$].

Figure 5 shows the $F0$ changes in median and range relative to plain speech. Both median and range varied significantly across speech styles [both $p<.001$]. All speech styles had a significantly higher median $F0$ than plain speech, with the exception of computer-directed speech which was statistically identical. Furthermore, in computer-directed speech, $F0$ range was significantly lower than in plain speech. In contrast, $F0$ range in infant-directed speech was significantly higher than that in plain speech.

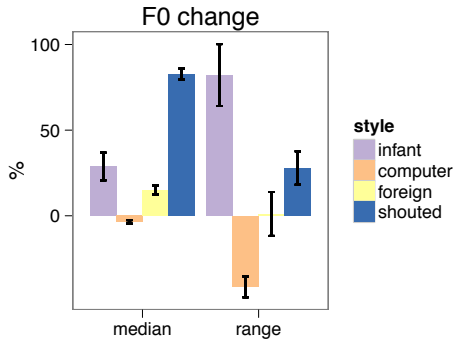


Figure 5: *Average changes of $F0$ median and range relative to plain speech across sentences.*

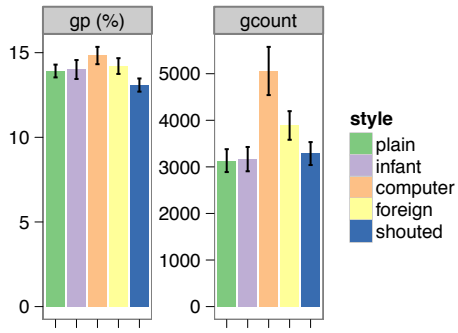


Figure 6: *Glimpse proportion and counts.*

Glimpse analyses (Figure 6) indicate that the proportion of speech which escaped masking was fairly similar across styles. The shouted style produced slightly fewer glimpses per second, lending further support to the idea that this was less a 'Lombard' response to noise—where increases in glimpse proportion are seen [22]—than an acted style. However, the count of glimpsed regions correlates well [$\rho = 0.94, p < .05$] with word identification rate. Glimpse counts are significantly higher

for computer-directed speech [$p < .001$] and somewhat higher for the foreigner-directed style [$p < .01$].

## 4. Discussion

The results of the current study suggest that much of the prosody-related intelligibility gain comes from durational increases, supporting those studies [3, 4, 7, 10] which found a durational advantage of clear speech. It is striking that infant-directed speech was less intelligible than plain speech, even though they shared many durational characteristics. For the infant-directed style, $F0$ range showed the clearest departure from plain speech. Without ruling out other modifications that might have been present in this acted speech style (e.g. expanded vowel space [23]; consonant hyper-articulation [24]), an extended pitch range may have been detrimental to speech intelligibility. Computer-directed speech – the most intelligible style – had significantly *reduced* pitch variation. It may be that a flat pitch countour helped listeners by providing continuity cues. These results contrast with earlier studies [12, 18], and suggest that the role of $F0$ in intelligibility may be more complex than previously posited.

The fact that glimpse count correlates well with intelligibility, while glimpse proportion does not, suggests that what matters to listeners in identifying sentences in noise where the amount of information is *fixed* (the same sentences were used in each speech style) is the total number of regions where the target speech escapes masking. Note that a strategy of adding short pauses, which may be beneficial in noise for other reasons, such as reducing cognitive load or clarifying word boundary locations, does not automatically increase the absolute glimpse count. On the other hand, segment elongation will typically result in more glimpses. Since computer-directed speech shows significant positive segmental duration changes across all segment types (Figure 4), it seems likely that these contributed to an increase in absolute glimpse count and perhaps was responsible for some of the reduction in WER observed for this style.

**Acknowledgement.** This study was supported by the LISTA Project FET-Open grant no. 256230 and by the SCALE Project under grant agreement 213850.

## 5. References

[1] R. M. Uchanski, "Clear speech," in *The Handbook of Speech Perception*, D. B. Pisoni and R. E. Remez, Eds. Oxford, UK: Blackwell, 2005, pp. 207–235.

[2] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *J. Acoust. Soc. Am.*, vol. 112, no. 1, pp. 272–284, 2002.

[3] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking Clearly for the Hard of Hearing II: Acoustic Characteristics of Clear and Conversational Speech," *J. Speech Hear. Res.*, vol. 29, no. 4, pp. 434–446, 1986.

[4] A. R. Bradlow, "Confluent talker- and listener-oriented forces in clear speech production," in *Laboratory Phonology*, C. Gussenhoven and N. Warner, Eds. Berlin: Mouton de Gruyter, 2003, pp. 241–273.

[5] A. Cutler and S. Butterfield, "Durational cues to word boundaries in clear speech," *Speech Comm.*, vol. 9, pp. 485–495, 1990.

[6] J. C. Krause and L. D. Braida, "Acoustic properties of naturally produced clear speech at normal speaking rates," *J. Acoust. Soc. Am.*, vol. 115, pp. 362–378, 2003.

[7] Z. S. Bond and T. J. Moore, "A note on the acoustic-phonetic characteristics of inadvertently clear speech," *Speech Comm.*, vol. 14, pp. 325–337, 1994.

[8] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, W. Hardcastle and A. Marchal, Eds. Dordrecht: Kluwer Academic Publishers, 1990, pp. 403–439.

[9] R. M. Uchanski, S. S. Choi, L. D. Braida, C. M. Reed, and N. I. Durlach, "Speaking Clearly for the Hard of Hearing IV: Further Studies of the Role of Speaking Rate," *J. Speech Hear. Res.*, vol. 39, no. 3, pp. 494–509, 1996.

[10] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *J. Acoust. Soc. Am.*, vol. 116, pp. 3108–3118, 2004.

[11] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Comm.*, vol. 51, pp. 1253–1262, 2009.

[12] J. Laures and K. Bunton, "Perceptual effects of a flattened fundamental frequency at the sentence level under different listening conditions," *J. Comm. Disord.*, vol. 36, pp. 449–464, 2003.

[13] R. M. Cox, G. C. Alexander, and C. Gilmore, "Intelligibility of Average Talkers in Typical Listening Environments," *J. Acoust. Soc. Am.*, vol. 81, pp. 1598–1608, 1987.

[14] A. R. Bradlow, G. Torretta, and D. B. Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Comm.*, vol. 20, no. 3, pp. 255–272, 1996.

[15] Y. Nejime and B. C. J. Moore, "Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 572–576, 1998.

[16] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing III: An attempt to determine the contribution of speaking rate to differences in intelligibility between clear and conversational speech," *J. Speech Hear. Res.*, vol. 32, no. 3, pp. 600–603, 1989.

[17] J. C. Krause and L. D. Braida, "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2165–2172, 2002.

[18] P. J. Watson and R. S. Schlauch, "The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours," *Am. J. Speech Lang. Pathol.*, vol. 17, no. 4, pp. 348–355, 2008.

[19] J. C. Krause, "Properties of naturally produced clear speech at normal rates and implications for intelligibility enhancement," Ph.D. dissertation, MIT, Cambridge, MA, 2001.

[20] W. A. Dreschler, H. Herschuure, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing aid assessment," *Audiology*, vol. 40, pp. 148–157, 2001.

[21] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, 2006.

[22] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Am.*, vol. 124, pp. 3261–3275, 2008.

[23] P. Kuhl, J. E. Andruski, L. A. Chistovich, E. Kozhevnikova, V. Ryskina, E. Stolyarova, U. Sundberg, and F. Lacerda, "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, no. 5326, pp. 684–686, 1997.

[24] A. Cristià, "Phonetic enhancement of sibilants in infant-directed speech," *J. Acoust. Soc. Am.*, vol. 128, no. 1, pp. 424–434, 2010.