Maximising objective speech intelligibility by local f_0 modulation

Julián Villegas[1,2]*, Martin Cooke*[1,2]

[1]Ikerbasque (Basque Foundation for Science), Spain
[2]Language and Speech Laboratory, University of the Basque Country, Spain
j.villegas@laslab.org,m.cooke@ikerbasque.org

Abstract

We investigated the effect on objective speech intelligibility of scaling the fundamental frequency (f_0) of voiced regions in a set of utterances. The frequency scaling was driven by maximising the glimpse proportion in voiced epochs, inspired by musical consonance maximisation techniques. Results show that depending on the energetic masker and the signal to noise ratio, f_0 modifications increased the mean glimpse proportion by up to $15\,\%$. On average, lower mean f_0 changes resulted in greater glimpse proportions. It was also found that the glimpse proportion could be a good predictor of music consonance.
Index Terms: roughness, glimpse proportion, objective speech intelligibility, musical consonance, fundamental frequency.

1. Introduction

Intelligibility of speech in noise can be improved by increasing the level of the signal [1] but that might not be practical or desirable since such increases could reach harmful levels or lead to listener discomfort, stress, or other noise-related disorders after long exposure. However, there is evidence that changes in other acoustic attributes correlate with intelligibility improvements. For example, speech produced in the presence of noise (so called 'Lombard speech') can be acoustically characterised, among many other factors, by increments in average f_0 and sound level, and decrements in speech rate and spectral tilt [2]; such a speech style is more intelligible than 'normal' speech when SNR differences between the two styles have been eliminated [3].

Decreasing spectral tilt has been found to greatly affect intelligibility [4, 5], while increasing speech rate seems to be detrimental [6]. Effects of f_0 modifications are somewhat less clear. Artificial modifications to the f_0 contour (specifically flattening, increasing f_0 range, inverting, and sinusoidally frequency modulation) have a deleterious effect on intelligibility of speech in noise [7], but when the masker is a competing talker, increasing the f_0 difference between two speakers was found beneficial [8]. Furthermore, whereas increasing average f_0 correlates with Lombard speech [3, 6], it has been found that it correlates to speech intelligibility improvements only for female speakers [9], has no significant correlation [10], or even has negative effects [11].

In this study, we investigate whether modifications of the f_0 contour can produce objective speech intelligibility increases when the changes are made locally, i.e., independently changing average f_0 in each voiced region of an utterance. Local changes are driven by a maximisation of the glimpse proportion [12] in a procedure derived from musical consonance maximisation.

Musical consonance (a complex concept involving sensory and non-sensory components) has been associated with a psychoacoustic percept called roughness [13]. In general terms, the rougher a sound mix is, the less consonant it is perceived [14]. We hypothesise that given the resemblance of voiced speech with some musical timbres and the apparent importance of the temporal envelope of a sound to decode speech [15], techniques to enhance musical consonance by 'softening' the temporal envelope of a sound mixture, could be applied in the speech realm to improve intelligibility as suggested by [16].

2. Roughness

Roughness is an auditory sensation associated with rapid amplitude variations of a sound signal [17]. Temporal envelope changes can be perceived as 'fluctuation strength' or 'roughness,' depending upon the modulation rate. Fluctuation strength has a maximum at about $4\,$Hz, and roughness peaks when the modulation rate is about $70\,$Hz [18]. Roughness tends to disappear when the modulation frequency is greater than $300\,$Hz. The continuous transition between the two percepts is at about 15–20 Hz. Roughness seems to be related to the auditory system's inability to accurately track more than one tone present in a single auditory filter [19].

Roughness has been linked to several auditory phenomena including voice quality, annoyance, and musical consonance [20]. In the latter, it has been shown that regardless of timbre, musical intervals judged as consonant correlate with local minima of roughness as a function of f_0 ratio [14].

Several models to compute roughness have been proposed, e.g., [21, 20]. One of the main differences between them is the way they account for envelope variations in each critical band. A detailed description of these models is beyond the scope of this article, but to illustrate the differences, whereas some approaches [21] use the spectral components of a sound to derive roughness, most models are based on the temporal variation in each auditory filter.

Based upon roughness minimisation, several "consonance enhancers" (e.g., [22]) have been proposed. These algorithms scale the f_0 of individual audio streams so the resulting mix has lower roughness than the original.

Glimpse proportion (GP) is a model-based quantification of the spectro-temporal regions where a signal escapes from energetic masking. It has been used to objectively assess the intelligibility of speech and is a reasonable predictor of subjective intelligibility [23, 24]. Although according to the definition of roughness GP seems to be related to it (i.e., when there is energetic masking, there maybe more than a single frequency component located in an auditory filter increasing causing roughness), a direct comparison between them is difficult since roughness is a psychoacoustic measure whereas GP is an objective measure. An indirect comparison, however, can be done by computing the GP of all frequency ratios within an octave as illustrated in Figure 1. Each line in this figure was computed

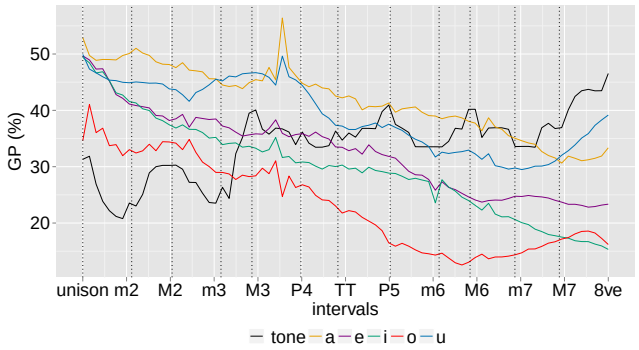


Figure 1: *Glimpse proportion of a synthetic duet. The local maxima of the complex tone correlates with musical intervals regarded as consonant. Dotted lines show pure intervals.*

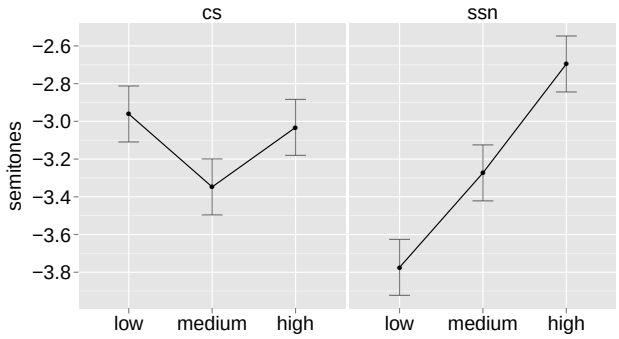


Figure 2: *Median f_0 modifications for each masker relative to the original utterances, discriminated by* SNR *level. The original median f_0 for all sentences was* 104 *Hz. Error bars indicate Fisher's least significant difference.*

using two identical signals with f_0= 120 Hz. While the f_0 of one signal was kept fixed, the other one was multiplied by a factor α which took 1200 values in the range of $1-2$. The signals were either vowels synthesised with Praat [25] with two formants at the suggested frequencies in the 'VowelEditor' or a complex tone comprising ten spectral components (fundamental and 9 harmonics) of the same magnitude and phase. All synthetic signals lasted one second and were sampled at 16 kHz. GP was evaluated between $60-7000$ Hz. In this figure, dotted vertical lines correspond to pure intervals, i.e., intervals that can be described by simple f_0 ratios such as $3/2$ (P5: a perfect fifth). These intervals have been considered the most consonant for harmonic timbres when they are heard in isolation of other intervals [26].

For synthetic vowels, GP scores decay with the size of the f_0 difference (in agreement with the trend of Figure 3(b)), having a prominent peak at about 4.5 semitones. For the harmonic tone, local maxima correlate surprisingly well with consonant intervals, e.g., the perfect fifth (P5), major third (M3), etc. Results of this idealised musical tone contrast with those of the synthetic vowels, for which certain commonalities among them were found but modifications in f_0 affect them somewhat differently. The same phenomenon has been observed in real instruments where different timbres have different consonance curves even when their spectra is mainly harmonic [14] suggesting that formants (related to 'timbre' in music) could play an important role in the effectiveness of f_0 modifications on intelligibility.

3. GP maximisation

3.1. Material

This study used 120 phonetically-balanced sentences from the Harvard lists [27], uttered by a male British-native speaker and recorded in a hemi-anechoic chamber at the University of Edinburgh. The sentences were down-sampled to 16 kHz and were mixed with one of two maskers: speech-shaped noise (SSN) with the same long term average spectrum of the male speaker and a female competing talker recorded under the same conditions. The utterances produced by the female speaker were semantically unrelated to the Harvard sentences. In the mix, each target sentence was preceded and followed by 500 ms of the energetic masker. This design eases a future comparison with subjective scores.

Three different SNRS per masker were used for mixing the signal and the masker. In a pilot study with English native listeners, SNRS of $1, -4, -9$ and $-7, -14, -21$ dB were found to have high, medium, and low intelligibility for the SSN and the competing taker maskers, producing approximately 75%, 50% and 25% words correct respectively.

3.2. Procedure

Voiced regions in each utterance were detected using Praat. The median f_0 and the baseline objective intelligibility (GP) were estimated over these regions for later analysis.

For each voiced region, 120 candidate intonations (each adjacent pair of candidates separated in frequency by the same ratio) were found by multiplying the f_0 contour by a factor α between $0.71-1.41$ (effectively an octave, centred at the region median f_0). The selected alternative intonation was chosen so it yielded the greatest glimpse proportion among the candidates. The glimpse proportion was evaluated on a narrower range ($60-750$ Hz) where resolved harmonics are mostly found.

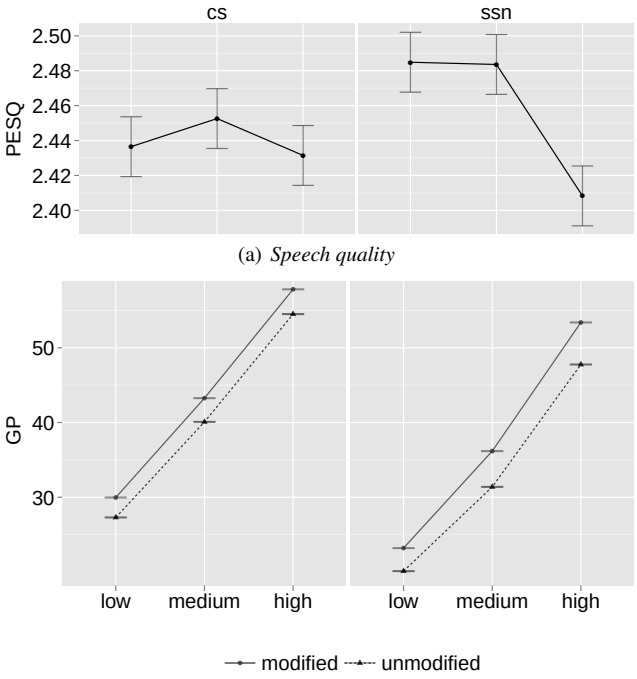
Finally, the modified voiced regions were used to compute the resulting median f_0 and GP, and were mixed with the unvoiced regions of the original utterance to form the modified sentences. Original and modified sentences were used to measure the degradation on speech quality using PESQ [28].

3.3. Results

The mean across median f_0 of the original utterances was 104 Hz. Optimisation of the GP metric acted to lower fundamental frequencies for all modifications as illustrated in Figure 2. The pattern of results was similar for both maskers, though the reduction was greater for SSN than CS at the lower SNR. Significant f_0 differences between SNRS were found in the SSN masker case, whereas only f_0 for medium SNR was significantly different in the competing speaker case. These results were confirmed with a three-way mixed-model ANOVA using masker (CS, SSN), SNR (low, medium, high), and treatment (with and without f_0 modification) as fixed effects. A main effect of treatment was found ($F(1, 119) = 1709.1, p < .001$).

Objective intelligibility and quality differences between modifications and original sentences are presented in Figure 3. In every condition, while objective intelligibility improved, objective speech quality was reduced. Table 1 summarise the relative mean changes in each condition.

For each masker, objective intelligibility and quality improvements were assessed by means of two-way mixed-model



(a) *Speech quality*

(b) *Glimpse proportions* (GP)

Figure 3: *Mean glimpse and* PESQ *scores, discriminated by* SNR *level.* PESQ *values for unmodified speech are all* 4.5.

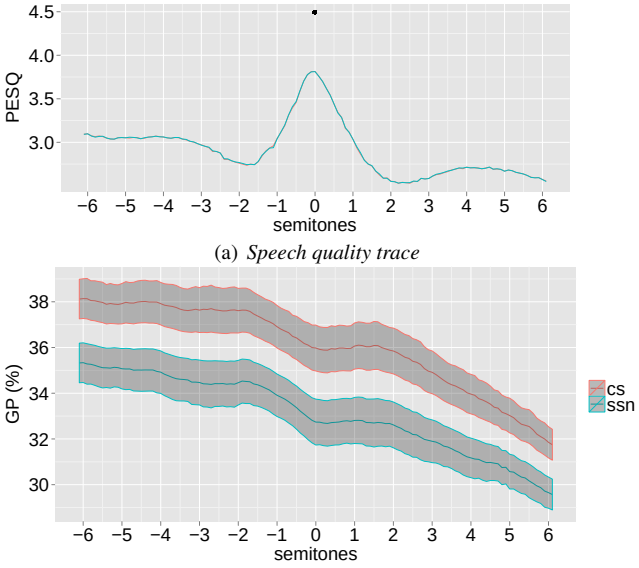
Table 1: *Changes in* GP *and* PESQ *of* f_0 *modifications as a percentage relative to the non-modified utterances.*

masker	SNR	GP	PESQ
speech-shaped noise (SSN)	low	15.3	−44.8
	mid	15.3	−44.8
	high	11.8	−46.5
competing speaker (CS)	low	9.8	−45.9
	mid	7.9	−45.5
	high	6.1	−46.6

ANOVA with SNR and treatment as fixed effects (with the same levels as before). Treatment in intelligibility was found significant [CS($F(1, 119) = 3041.5$, $p < .001$) and SSN ($F(1, 119) = 3761.3$, $p < .001$)] as well as in objective quality [CS($F(1, 119) = 96.85$, $p < .001$) and SSN ($F(1, 119) = 61.07$, $p < .001$)].

Regardless of energetic masker and SNR level, PESQ scores were greatly affected by f_0 modifications. As shown in Figure 4(a), quality for higher and lower pitches degraded rapidly in the vicinity of the median f_0 finding minima at about ± 2 semitones and growing from there just a fraction, in the case of negative f_0 changes, almost asymptotically at the score of 3.0. In case of positive changes, PESQ scores achieve a lower values, which suggests that positive pitch scalings are detrimental for GP and PESQ.

Part of the degradation can be attributed to artefacts in the resynthesis process: the PESQ score for the same signal used as unmodified and modified speech is always 4.5; scaling the signal with $\alpha = 1$ should results in no f_0 and PESQ score changes if the resynthesis process was transparent, however an average degradation of about 0.7 was found as shown in Figure 4(a). The traces in Figure 4(b) showthat lower pitches yielded in average higher GP scores.



(a) *Speech quality trace*

(b) *Objective intelligibility trace*

Figure 4: *Mean changes in* PESQ *and* GP *computed over an octave centred at the median* f_0 *of each utterance in the mid* SNR *condition.*

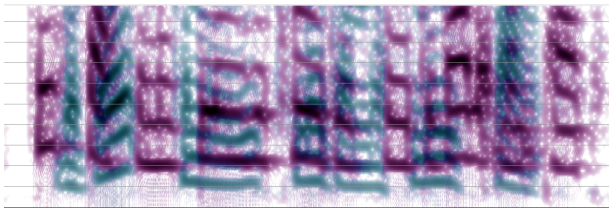
4. Discussion

One unanticipated finding was that lower median f_0 produced higher objective intelligibility scores. It could be that objective intelligibility is improved because of the larger f_0 differences between the signal and the masker (as found in [8]), or because of an increment in the number of harmonics within the same spectral region as suggested in [29]. To clarify this issue, the simulation was repeated using the female speech as target and the male speech as masker, and using the same (male) speaker as both target and masker. In these simulations, only the high intelligibility SNR for the competing speaker was used (−7 dB). It was assumed that if the f_0 difference was the leading cause, in case of the female speech as target, f_0 changes should be positive rather than negative.

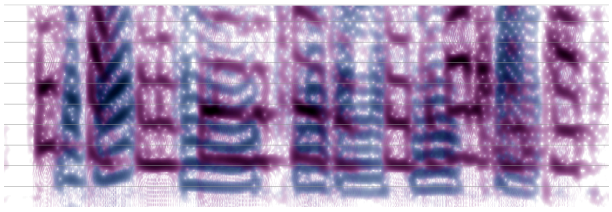
It was found that an average of −4.4 semitones produced relative increases of about 3 % in the case of female target with male masker (intelligibility baseline of 50 %), and an average of −3.7 semitones produced 15 % relative increases in the male target and masker case (intelligibility baseline of 22.6 %). These findings do not support the hypothesis that larger f_0 differences lead to objective intelligibility improvements. On the other hand, as shown in the overlaid spectrograms of Figure 5, some spectral components masked in the original speech (Figure 5(a)) are unmasked in the modified speech (Figure 5(b)), suggesting that the enhancements may be caused by an increase in the number of harmonics resulting from a lowering of median f_0 , which in turn increases the likelihood that some will escape masking (e.g., by harmonics from the competing talker). Fundamental frequency increases are a common correlate of Lombard speech so the results of these simulations are surprising. Given that the length of the vocal tract is fixed, producing low pitches while speaking louder may be difficult to achieve in practice, but nevertheless could be implemented in speech output technology.

5. Conclusions

At least objectively, maximising the glimpse proportion (GP) in the voiced part of utterances, a procedure inspired in musi-



(a) *original signal (blue) and the competing talker (purple)*



(b) *enhanced signal (blue) and the competing talker (purple)*

Figure 5: *Spectrograms (up to* $1\,kHz$*) of one sentence masked with the female* CS *mixed at* -7 SNR.

cal consonance enhancers, produces beneficial results. It was also shown that GP may be correlated with musical consonance. Contrary to the behaviour of speakers in noisy environments, lowering f_0 seems to be more beneficial for increasing objective intelligibility of speech with the selected maskers (SSN and female competing talker). The downwards change is not likely to be due to a maximisation of the f_0 difference between the signal and the masker but rather for an increase in the number of harmonics for a given spectral region. Finally, although GP has been found to be a good predictor of subjective intelligibility scores, the underlying model is based on energetic masking and does not account for other masking effects (e.g., informational, or resulting from cognitive loading) or indeed the role that f_0 seems to play in source separation [30], so these results need to be verified subjectively and with a larger sample of utterances and maskers.

Acknowledgements. The authors thank EU Future and Emerging Technology (FET–OPEN) Project LISTA (The Listening Talker) and staff at the Centre for Speech Technology Research of the University of Edinburgh for their collaboration in making available speech material.

6. References

[1] H. Lane and B. Tranel, "The Lombard Sign and the Role of Hearing in Speech," *J. Speech Hear. Res.*, vol. 14, no. 4, p. 677, 1971.

[2] J. Junqua, "The Lombard reflex and its role on human listeners and automatic speech recognizers," *J. Acoust. Soc. Am.*, vol. 93, pp. 510–524, 1993.

[3] W. V. Summers, D. Pisoni, R. Bernacki, R. Pedlow, and M. Stokes, "Effects of noise on speech production: Acoustic and perceptual analyses," *J. Acoust. Soc. Am.*, vol. 84, no. 3, pp. 917–928, 1988.

[4] 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, 2004.

[5] 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.

[6] V. Hazan and D. Markham, "Acoustic-phonetic correlates of talker intelligibility for adults and children," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3108–18, Nov. 2004.

[7] S. E. Miller, R. S. Schlauch, and P. J. Watson, "The effects of fundamental frequency contour manipulations on speech intelligibility in background noise," *J. Acoust. Soc. Am.*, vol. 128, no. 1, pp. 435–443, 2010.

[8] P. F. Assmann, "Fundamental frequency and the intelligibility of competing voices," in *Proc. 14 Int. Cong. of Phonetic Sciences*, 1999.

[9] J. Barker and M. Cooke, "Modelling speaker intelligibility in noise," *Speech Comm.*, vol. 49, pp. 402–417, 2007.

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

[11] J. H. Ryalls and P. Lieberman, "Fundamental frequency and vowel perception," *J. Acoust. Soc. Am.*, vol. 72, pp. 1631–1634, 1982.

[12] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, pp. 1562–1573, 2006.

[13] E. Terhardt, "Ein psychoakustisch begründetes Konzept der musikalischen Konsonanz (A Psychoacoustic-Driven Concept of Musical Consonance)," *Acustica*, vol. 36, pp. 121–137, 1976, in German.

[14] W. Sethares, *Tuning, Timbre, Spectrum, Scale*, 2nd ed. London: Springer, 2005.

[15] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, Mar. 2002.

[16] J. F. Culling and C. J. Darwin, "Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1559–1569, 1994.

[17] H. von Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, II english ed. Dover Publications, 1954, ch. VIII, On the beats of simple tones.

[18] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed., ser. Springer series in information sciences. Berlin: Springer, 2007.

[19] J. Beament, *How We Hear Music: The Relationship between Music and the Hearing Mechanism*. Rochester, NY: Boydell Press, 2001.

[20] P. Daniel and R. Weber, "Psychoacoustical Roughness: Implementation of an Optimized Model," *Acta Acustica United with Acustica*, vol. 83, pp. 113–123, 1997.

[21] R. Plomp and W. Levelt, "Tonal Consonance and Critical Bandwidth," *J. Acoust. Soc. Am.*, vol. 38, no. 4, pp. 548–560, 1965.

[22] J. Villegas and M. Cohen, "Roughness Minimization Through Automatic Intonation Adjustments," *J. of New Music Research*, vol. 39, no. 1, pp. 75–92, 2010.

[23] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified hmm-based synthetic speech in noise?" in *Interspeech*, 2011.

[24] Y. Tang and M. Cooke, "Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints," in *Interspeech*, 2011.

[25] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 2012, available [Mar. 2012] from www.praat.org.

[26] D. Benson, *Music: A Mathematical Offering*. Cambridge, UK: Cambridge University Press, 2008, ch. 4. Consonance and dissonance.

[27] E. H. Rothauser, W. D. Chapman, N. Guttman, H. R. Silbiger, M. H. L. Hecker, G. E. Urbanek, K. S. Nordby, and M. Weinstock, "IEEE recommended practice for speech quality measurements," IEEE *Trans. on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225–246, Sep. 1969.

[28] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, vol. 2, 2001, pp. 749–752.

[29] J. Chalupper, "Aural exciter and loudness maximizer: What's psychoacoustic about "psychoacoustic processors"?" in *Proc. 109 Audio Eng. Soc. Conv.*, 2000.

[30] C. J. Darwin, "Listening to speech in the presence of other sounds," *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 363, no. 1493, pp. 1011–1021, Mar. 2008.