Acoustic-phonetic dimensions of speaker intelligibility

Valerie Hazan and Duncan Markham

University College London, U.K.

E-mail: v.hazan@phon.ucl.ac.uk, dunx_2000@yahoo.com

ABSTRACT

Word intelligibility rates were obtained for a set of 45 speakers from a homogeneous accent group. 135 listeners were tested: 45 adults, 45 11-12 year olds and 45 7-8 year olds. Intelligibility in low-level noise varied significantly across speakers with mean word error rates ranging from 3.6% to 18.8%. Error rates were higher for the younger child listeners but the relative intelligibility of speakers across listener groups was highly consistent. Next, acoustic-phonetic measurements were made on the speaker database. Total energy in the 1-3 kHz region and mean word duration together predicted over 60% of the variability in the intelligibility data. However, correlations between intelligibility and acoustic-phonetic measures varied across speaker groups, with no correlations obtained for child speakers, and the profiles of the 'best' and 'worst' speakers highlighted the considerable diversity of factors contributing to intelligibility. These results confirm the difficulty in finding reliable acoustic-phonetic correlates of speaker intelligibility.

1. INTRODUCTION

The fact that some speakers are more intelligible than others is well attested but we do not yet fully understand what acoustic-phonetic characteristics correlate with intelligibility. Previous studies have found correlations between intelligibility and a number of acoustic-phonetic measures, including word and vowel duration, size of vowel space and cues to consonantal contrasts [1], fundamental frequency range, and precision of articulation [2]. There is also some evidence that female speakers may be more intelligible on average than male speakers [2]. However, the number of speakers used in these studies has typically been small and the acoustic-phonetic correlates of intelligibility are not consistent across studies.

There has also been insufficient attention to the issue of whether a speaker's 'inherent' intelligibility is primarily determined by the acoustic-phonetic characteristics of their speech or whether it may be related to listener characteristics (e.g. a listener's experience with particular types of voices). The evidence that detailed episodic information about a speaker's voice is retained in long-term memory [3] and that words produced by voices that were perceptually similar to previously-heard voices were better identified than words produced by dissimilar voices [4] suggests that a listener's experience of different speakers may impact on perceptual abilities. Given children's greater exposure to children and women's voices, it may be of particular interest to see if they find such voices more intelligible.

The aim of the study was therefore to examine whether adults and children varied in terms of which voices they found to be more intelligible, and whether they varied in their ability to normalise to different speakers. We also wished to investigate whether intelligibility was correlated to specific acoustic-phonetic characteristics.

2. PERCEPTION STUDY

2.1 STIMULI

The UCL Markham test, a new word-level test, was developed: requirements were that it would be appropriate for testing children aged seven and above, would maximise the number of consonant confusions and could be used in an open-set response format. The test consists of 124 key-words that are familiar to 7 year olds and that adequately cover all frequent consonant confusions [5].

2.2 SPEAKERS AND LISTENERS

The speaker population consisted of 45 speakers of British English with a neutral or mild south-eastern English accent: 18 women (mean age: 33;11 yrs), 15 men (mean age: 30;7 yrs), six girls and six boys (mean age: 13;2 yrs).

The listener population consisted of 135 listeners: 45 children aged 7-8 years, 45 children aged 11-12 years and 45 adults. Listeners were only included if they had pure tone thresholds of less than 25 dB HL at octave intervals from 0.5 to 8 kHz. The CELF Recalling Sentences test was also presented as a screening of language ability. All listeners completed this test within criterion.

2.3 PROCEDURES

Speech recordings, made in an anechoic chamber, were recorded to DAT at a sampling rate of 44.1 kHz and a simultaneous recording of laryngeal activity was made using a Laryngograph [6]. The recordings were transferred to PC at the original sampling rate, and segmented into individual files. Each of the stimuli was leveled to a fixed RMS level and 20-speaker babble (MRC-IHR) was added to produce a SNR of +6 dB, in order to avoid ceiling effects in the perceptual task.

In order to evaluate the effect on intelligibility of 'normalising' information, listeners were tested in two conditions. In the 'triplet' condition, a carrier phrase ('the next three words are' *or* 'and now please say') was followed by a set of three key-words produced by the same speaker, each separated by a 200 ms gap whilst the 'single-word' task involved the presentation of individual words without a carrier sentence. In the triplet condition,

ISBN 1-876346-48-5 © 2003 UAB

each listener heard 25 key -words from each of 15 speakers, presented in a fully randomised fashion. In the single-word condition, each listener heard 27 key-words from each of 15 new speakers, also presented in a fully randomised fashion.

2.4 Test methodology

Listeners were tested individually in a quiet room at their school or university. The test was computer-controlled, with stimuli presented via headphones at a fixed comfortable listening level. Stimulus presentation was timed manually, each word or triplet being initiated as soon as the previous key-word(s) had been repeated by the listener. The experimenter who was a trained phonetician transcribed erroneous responses by hand. At the first testing session, the hearing screening test was carried out and listeners heard 72 triplets and 150 words. At the second, after an interval of approximately one week, listeners heard 63 triplets and 225 words and completed the language screening test. Total testing time was around 60 minutes.

2.5 Results

A mean word error rate per speaker group (women, men, boys, girls) was obtained for each of the 135 listeners, by averaging error rates obtained for speakers from each speaker group in both presentation conditions (see Figure 1). A repeated-measures analysis of variance (ANOVA) showed that the effect of speaker group was significant [F(3, 396)=18.75; p<0.0001)]; pairwise comparisons showed that error rates for women were lower (8.3%) than error rates for other speaker groups (ranging from 9.8 to 10.3%). Children were not significantly less intelligible than men but there was greater variance in error rates for all listener groups when listening to children than to adults. The effect of listener group was also significant [F(2, 132)=13.67; p<0.0001]: 7-8 year old listeners made more errors (10.9%) than 11-12 year olds (9.3%) and adults (8.8%), but there was no difference between the older children and adults. There was no significant speaker group by listener group interaction so no evidence that listeners find speakers of their own sex or age to be more intelligible. The higher intelligibility of women speakers was consistent across listener groups (see Figure 1).

The next analysis focused on the effect of test condition (triplet vs. single word). A mean error score for triplet and word conditions calculated over all speakers was obtained for each listener. A repeated-measures ANOVA showed that the word error rate in the triplet condition did not differ significantly from that in the single-word condition; there was no significant listener group by condition interaction. None of the listener groups therefore made fewer errors when speaker information was provided via a precursor sentence.

Despite the consistent difference in word error rates between younger children and other listeners, error rates for individual speakers were very strongly correlated across listener groups (e.g. adult-OC r=0.952; adult-YC r= 0.900). Relative speaker intelligibility is therefore consistent across adult and child listeners.

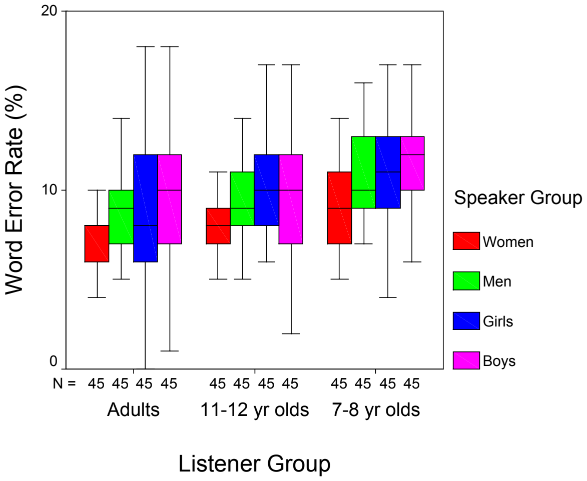

Figure 1: Box plots showing the word error rates by the three listener groups (adults, 11-12 yr olds, 7-8 yr olds) for the four speaker groups (women, men, girls, boys).

3. ACOUSTIC-PHONETIC MEASURES

3.1 Methodology

Long Term Average Spectrum (LTAS). A long-term average spectrum (LTAS) was calculated for each speaker from a file containing all 124 key-words normalised for level. Analyses used an FFT with a length of 2048 sample points, and windows overlapping by 1024 points, giving a value for the LTAS at multiples of 21.53 Hz. These spectra were then smoothed with a two-octave wide Hamming window over the frequency range 50 Hz-10 kHz. Two measures of LTAS were obtained: spectrum slope was derived by fitting a curve between 0.5 and 4 kHz, and the total energy in the 1 to 3 kHz region was also calculated (LTAS 1 – 3 kHz).

Fundamental Frequency. Measures of fundamental frequency (F0) were obtained by analysing the laryngograph signal obtained during the reading of a two-minute text. Four measures were examined: F0 median, F0 range (in octaves), closed-phase ratio median and F0 irregularity.

Duration. Duration measurements were made on a subset of 41 key-words from each of the 45 speakers (n=1845) that had been segmented and phonetically annotated. This subset was chosen to contain sets of words differing in a single phoneme (e.g. cheap, cheat, cheek) and to contain a representative range in terms of consonant manner and place of articulation. Mean word duration per speaker was calculated as a general measure of speaking rate.

CV amplitude ratio. CV ratio measurements were made on the same subset of 41 key-words per speaker. First, rms energy was calculated using a 1ms window, then average intensities were calculated automatically for each phonetically-annotated segment. For each speaker, CV ratios for plosive, fricative and nasal consonants were made on measurements of between 12 and 14 tokens for each manner of articulation.

Vowel formants. Words were selected that contained the vowels /ae/ (9 tokens), /i/ (8 tokens) , and /u/ (4 tokens). First and second formant frequencies were measured manually in the steady-state region of the vowel, using a simultaneous display of the speech waveform, spectrogram and spectral cross-section. The frequency measures obtained were transformed to the erb scale, an auditory frequency scale [7]. The following measures were then derived as measures of vowel space: the Euclidian distance between F1 and F2 for each vowel, the difference in F1 between /i/ and /ae/ and the difference in F2 between /i/ and /u/.

3.2 Results[1]

Long-term average spectrum (LTAS). LTAS slope varied from –7.86 to –13.3 dB/decade (mean: -10.3, s.d. 1.47). The total energy in the 1-3 kHz region (LTAS 1-3 kHz) was –1.93 dB (s.d. 2.88) for women (n=18), -3.36 dB (s.d. 2.72) for men (n=15) and –1.40 dB (s.d. 1.81) for children (n=12). An ANOVA showed that there was no significant effect of speaker group and of speaker sex on these two LTAS measures. There were weak or no correlations between mean word error rate and LTAS slope. However, relatively strong correlations were obtained between word error rate and the LTAS (1-3 kHz) measure when carried out on the whole speaker set (r=-0.635). When carried out on separate speaker groups though, the strongest correlation was obtained for men (r=-0.752), and with no significant correlation for child speakers.

Fundamental frequency. As expected, the effect of speaker group on F0 median was significant [F(3,34)=47.64; p<0.0001]: men had a lower F0 median than all other groups, but there were no differences between women and children. There were no significant effects of speaker group or sex on F0 irregularity, F0 range or closed-phase ratio. The only significant, if weak, correlations were between the F0 irregularity measures and word error scores for older child listeners (r=0.38; p=0.02) and younger children listeners (r=0.35; p=0.03).

Word duration. Mean word duration was 0.512 sec (s.d. 0.05) for women, 0.453 sec (s.d. 0.04) for men and 0.503 sec (s.d. 0.04) for children; the effect of speaker group on mean word duration was significant [F(2,42)= 8.248; p=0.001]: men had a faster speaking rate than women and children. Overall, mean word duration and word error rates were significantly correlated (r=-0.358, p<0.02). When calculated separately for each speaker group, correlations were found to be relatively strong for adult male speakers (r=-0.645, p=0.009) but non-significant for women and child speakers (See Figure 3).

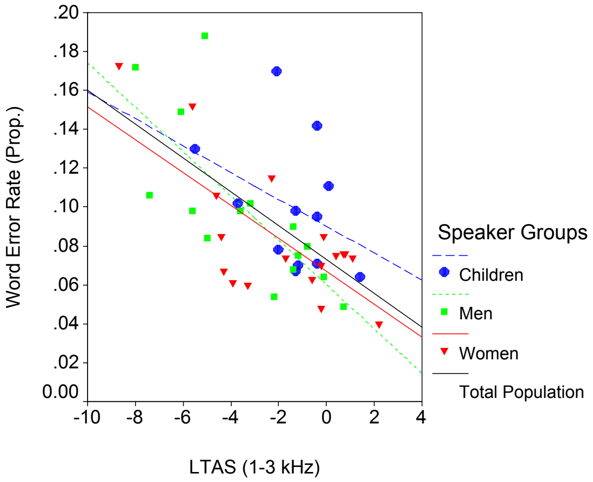

Figure 2: Correlations between total energy in the 1-3 kHz region (dB) (LTAS 1-3 kHz) and word error rates for each speaker group (children, men and women).

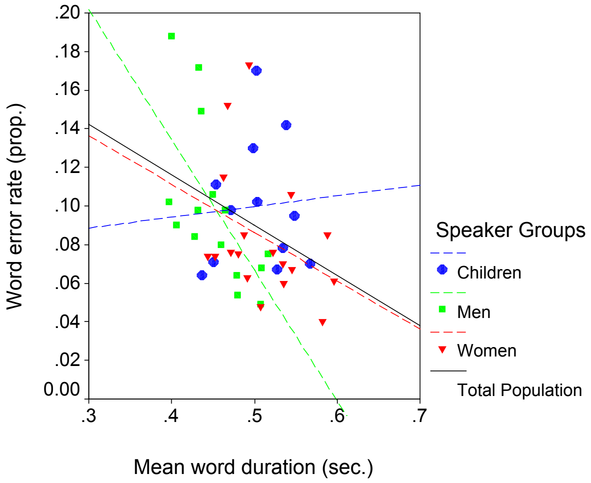

Figure 3: Correlations between mean word duration (reflecting speaking rate) and word error rates for each speaker group (children, men and women).

CV amplitude ratios. The nasal/vowel ratio was calculated for word-initial and word-final nasals; the effect of speaker group was significant [F(2, 42)=8.009, p<0.001]: nasal/vowel ratio in men's speech was lower than in women's or children's. Significantly lower nasal/vowel ratios were obtained in initial position (-0.274 dB) than in final position (-7.06 dB). There were no significant correlations between nasal/vowel ratio and mean word error rates. The fricative/vowel ratio was calculated based on 9 tokens per speaker. The effect of speaker group was weakly significant [F(2,42)=4.045; p<0.05] with a weaker CV ratio obtained in children's speech than women's. There were no significant correlations between mean word error rates and CV ratio for fricatives. Stop-burst/vowel ratios were calculated for 23 tokens per speaker. Final plosives had a significantly lower CV ratio than initial plosives [F(1,42)=96.25; p=0.0001] but there was no significant effect of speaker group. No significant correlations were

[1] The data for boys and girls was grouped in all these analyses (women n=18; men n=15; children n=12)

found between stop/vowel ratios and word error rates.

Vowel Formant measures. Mean values of F1 and F2 (expressed in erb) and Euclidian distance between F1 and F2 were calculated for the three vowels /i/, /ae/ and /u/ for all speaker groups. Here, separate means were calculated for boys and girls in order to evaluate any difference in formant values between these two groups. Repeated-measures ANOVAs were carried out on mean Euclidian distance calculated per vowel and per speaker. The effect of vowel was significant [F(2,82)=430.48; hp2=0.913; p<0.001], as expected and this factor alone accounts for 91% of the overall variance. The effect of speaker group was also significant [F(3,41)=37.98, hp2=0.735; p<0.001] and post-hoc tests (Tukey's HSD) showed that men differed from the other three groups of speakers but that there was no difference in mean Euclidian distance for /i/, /ae/, /u/ between women, boys and girls. The only significant correlation was between word intelligibility and the difference in F2 value between /i/ and /u/ (r=0.349; p<0.05). When groups were analysed separately, correlations between formant values and word intelligibility were relatively strong for adult male speakers but weak for child speakers.

Regression analyses. A multiple regression analysis was applied to the data using a forward stepwise method. Measures that had been found to be correlated with intelligibility for one or more speaker groups (LTAS 1-3 K, word duration, F0 irregularity, difference in F2 between /i/ and /u/) were included in the analysis. The dependent variable was the word intelligibility rate aggregated over all listener groups for all speakers. The final model included only two metrics: LTAS 1-3K and word duration. The R^2 for the model with LTAS 1-3 kHz and word duration measures as predictors was 0.61 with an R^2 of 0.45 for the long-term average spectrum measure alone and a change in R^2 of 0.16 when the word duration metric was added.

4. CONCLUSIONS

This study involving a much range of both speakers and listeners than previous studies of speaker intelligibility confirmed that women were on average more intelligible than men. However, this effect was much less clear-cut than in a previous study of speaker intelligibility [2]. Despite the higher intra-speaker variability in speech production typically seen in children below the age of 14 [8], child speakers were, as a group, no less intelligible than adult speakers, either for adult or child listeners. This study also confirmed that acoustic-phonetic characteristics of a speaker’s utterance are the prime determinant of speaker intelligibility, given that relative speaker intelligibility was highly correlated across groups of listeners differing in age and sex.

Defining what these acoustic-phonetic characteristics might be is a difficult task, given the differences in patterns of correlations for different speaker groups, and also given the variability that is seen in groups of ‘good’ or ‘poor’ speakers. In our study, speaking rate and amount of energy in the 1-3 kHz frequency range, which is rich in acoustic cue information, were shown to have the highest correlation with word error rates, but these factors are not consistent across studies. Even in our data, although these factors do appear to promote high intelligibility, there are neither necessary nor sufficient to produce naturally-clear speech. Some speakers in our study were highly intelligible because of acoustic-phonetic characteristics other than those measured, as they did not rank high on any of the spectral, intensity or durational measures made. These results confirm the difficulty in finding reliable acoustic-phonetic correlates of speaker intelligibility.

ACKNOWLEDGEMENTS

Funded by the Wellcome Trust (055651/ Z/98/ JRS/ JP/ JAT). We thank staff and pupils at: City of London School for Boys, St Christopher's School, Notting Hill and Ealing Secondary School, Broomwood Hall, Newton Preparatory School, St Paul's Cathedral School, Thomas's Preparatory School, Battersea.

REFERENCES

[1] Z.S. Bond, Z.S. and T.J. Moore, “A note on the acoustic-phonetic characteristics of inadvertently clear speech,” *Speech Com*., vol. 14, pp. 325-337, 1994.

[2] A.R. Bradlow, G.M. Torretta, D.B. Pisoni., “Intelligibility of normal speech .1. global and fine-grained acoustic-phonetic talker characteristics,” *Speech Com*., vol. 20, pp. 255-272, 1996.

[3] T.J. Palmeri, S.D. Goldinger, & D.B. Pisoni, “Episodic encoding of voice attributes and recognition memory for spoken words”. *Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 19, pp.* 309-328, 1993.

[4] S.D. Goldinger “Words and voices: Episodic traces in spoken word identification and recognition memory”. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, pp. 1166-1183, 1996.

[5] D. Markham and V. Hazan, “The UCL Speaker Database”, *Speech, Hearing and Language: UCL Work in Progress*, vol. 14, pp. 1-17, 2002.

[6] A.J. Fourcin, “Electrolaryngographic assessment of vocal fold function,” *J.Phon*., vol. 14, pp. 435-442, 1982.

[7] B.R. Glasberg and B.C.J. Moore, “Derivation of auditory filter shapes from notched-noise data”, *Hearing Res*. vol. 47, pp. 103-138, 1990

[8] S. Lee, A. Potamianos, S. Narayanan , “ Acoustics of children's speech: developmental changes of temporal and spectral parameters,” *J. Acoust. Soc. Am***.** , vol. 105, pp. 1455-1468, 1999.