INTELLIGIBILITY OF SUNG WORDS IN POLYTEXTUAL SETTINGS

Sarah Hawkins[1], Kate Honey[1], Sarah Knight[2], Antje Heinrich[2]

[1]University of Cambridge, [2]MRC Institute of Hearing Research, Nottingham
sh110@cam.ac.uk

ABSTRACT

Three experiments used word-spotting to examine influences of phonetic and musical parameters on intelligibility of closed-set but unpredictable words in polytextual singing. Main comparisons were: 3 musical genres (medieval polyphonic motet, similar but homophonic motet, jingle); harmony (consonant, dissonant); keyword phonetic properties (‘acoustic contrast’, vowel length, vowel quality); contrast between keyword and competitor word (Onset, Vowel, Coda). Results showed strong effects of phonetic and musical parameters that affect pitch continuity and rhythm. Vowel quality affected responses but with no discernibly consistent pattern.

Keywords: polytextual singing; intelligibility

1. INTRODUCTION

The intelligibility of sung words is important not just to composers, performers and listeners, but also as a testbed for exploring speech intelligibility in adverse listening conditions, especially since, in ensemble singing, ‘background noise’ is intentionally intrinsic to the signal, and may contribute to the message.

Sung text can vary as much or more than natural speech: in genre, and in rate, melodic, harmonic and rhythmic properties within genre. Yet singing adds extra constraints on intelligibility. In addition to high pitch necessarily reducing formant definition and high vowels being less intelligible when f0 is higher than F1 (concert A=440 Hz; sopranos’ f0 is often higher), Western classical singing deliberately reduces vowel contrast to preserve equal loudness and timbre (voice quality) across a wide pitch range, and may reduce formant definition by avoiding placing vocal-tract resonances at multiples of f0 [3]. Clarity of articulation can enhance intelligibility, but the genre and music itself can impose limits [14].

Beyond these sorts of facts, little is known about what factors intrinsic to the signal influence sung word intelligibility. The few studies in the literature mainly focus on isolated vowels sung in quiet [1, 7, 9] or words in carrier phrases [4, 12]. They confirm intelligibility losses due to vowel centralization, and to consonant errors that reflect acoustic properties of segments, music, and probably their interaction, and phonotactic and lexical knowledge. The cited work was conducted with simple stimuli under controlled conditions. In a different approach, [8] showed that key factors affecting spoken word intelligibility also affect their intelligibility when sung by one voice in novel sentences in live concerts: intelligibility was higher when words were more predictable, had good signal-to-noise (SNR) ratios, and least low-frequency masking, though there were some strong interactions, including with listener characteristics. Parallels with spoken word intelligibility are also reported for number of hearings, meaningfulness and listener experience in solo and 3-voice unison [5, 6].

In much singing, however, the ‘target’ text is neither solo nor the only text being sung, nor is it even clear that word intelligibility is important. Medieval motets offer a prime example. Typically polyphonic (each voice is melodically independent) and polytextual (each voice sings different words), it was not infrequent for a motet to have three voices in three languages—French, English, Latin. Some motets may have been intended to be enjoyed from reading rather than singing. Moreover, though the style of medieval singing is not known, today’s performers favour a vocal timbre or ‘choral blend’ which makes it hard to distinguish separate voices. Yet at times they want the words to be understood.

This type of singing motivated the present work. Three questions were asked. How intelligible is such polytextual polyphonic singing? Do phonetic factors affect its intelligibility? Does harmonisation affect intelligibility? In a word-spotting task 3 experiments addressed these questions for motets sung normally (E1); with target voice at a favourable SNR (E2); for consonant vs dissonant harmony without polyphony, and for a dissonant vs solo jazzy/jingle setting (E3). All Methods are presented first, and then Results.

2. GENERAL METHODS

2.1 Materials

The first experiment addressed medieval motet singing. The music written for it shaped the verbal material (lyrics) used in all experiments. The full verbal dataset comprised 24 keywords, each in 2 target sentences, sung as the top voice, Alto (A). Each keyword was sung against each of three competitor words, placed in a different sentence from the target sentence, and sung as the second voice (Tenor (T)). Most experiments included a hummed third voice (Baritone (B)). Thus the typical stimuli were polytextual, with two competing texts above a third voice hummed on [ŋ]. Sentences were semantically anomalous. Words were monosyllabic and there were few function words. Thus all words were strongly unpredictable.

2.1.1 Verbal material ('lyrics')

The 24 keywords were monosyllabic animal words, Celex mean wordform spoken frequency 2.4 (sd 3.3) written 7.7 (sd 9.5); mean neighbourhood counts of 25.2 (sd. 13.6, Wordmine 2). Each keyword was to be heard against 3 competitor words which contrasted with the keyword's Onset (Odiff), Vowel (Vdiff) or Coda (Cdiff). E.g. keyword *sheep* had competitors *keep, sharp, sheaf* for O, V and C contrasts, none being animal words. Each keyword was placed towards the middle of its semantically-anomalous sentence. Some of each type of sentence contained other animal words as foils to ensure listeners attended throughout. Example:

Keyword: Pine to my soul thank fuse **lamb** hatch by make or low so then to haul pig theme bill yen for cane.

Odiff: Mass knack feel hag or whim **dam** tore with goal as teak tease more wake loss caught in ease.

Vdiff: Mass knack feel hag or whim **limb** …[as for Odiff]

Cdiff: Mass knack feel hag or whim **lass** …[as for Odiff]

Keywords varied three phonetic parameters: Acoustic Contrast (AC, high vs low), Vowel Length (VL, long vs short), and Vowel Quality (VQ, front vs low/central vs back). AC, a parameter devised by the first author, was intended to contrast degree of spectral change. Low AC words mainly had voiced consonants, and relatively little spectral change e.g. *bee worm mole*. High AC words typically had voiceless obstruents, causing f0 discontinuities at word boundaries, and more featural variety e.g. *sheep cat stoat*. VL was tense vs lax. VQ categories were imperfect due to animal word availability: front [i eɪ ɪ ɛ], low/central [ɑ ɜ a], back [u ɔ ɒ ʌ əʊ].

There were 2 keywords for each of these 2AC x 2VL x 3VQ = 12 phonetic conditions. Each keyword and its three competitors appeared in two sentence settings, totalling 24 keywords x 2 settings x 3 competitors = 144 stimuli.

2.1.2 Music

Four 3-part extracts were adapted from the 14th century *Roman de Fauvel.* They had 2-3 phrases. Three had 7 bars and one had 6. All had 3/2 time signature, legato conjunct melodies, modal tonality, and lay within a twelfth (E3 - B4). Keywords were on-beat, in a bar with no syncopation, part-crossing, unison, big pitch leaps, melisma, sustained notes in a single voice, chromaticism or ornamentation. Each keyword was heard in two different music extracts.

Stimuli were recorded in a studio by experienced motet singers, from a microphone about 2 m from the singers, and from 3 head-mounted close-talking mikes. 3 musicians listened for consistency and quality. Unsatisfactory stimuli were re-recorded.

2.2 Procedure (Listening Task)

Listeners, tested individually, sat at a computer screen and were asked to type in animal words they heard in the top voice, in real time. They first saw all 24 keywords on the screen, and heard practice items until they felt confident. During the test, keywords were to hand at all times. Each session ended with a computerised demographic questionnaire.

3. SPECIFIC EXPERIMENTAL METHODS

3.1 E1, P: Polyphonic motet, classic blend

Stimuli were rotated across extracts in a 6-group nested design such that each listener heard all music extracts and conditions, but each keyword once only i.e. with only 1 music extract and 1 competitor word.

60 native-speakers of English, no history of speech or hearing disorders, and normal or corrected-to-normal vision, heard all 144 stimuli, with standard choral blend from the single microphone. These 60, plus 60 others from E2, were aged 17-35 years, mean 24.6 years.

3.2 E2, P+SNR: Polyphonic motet, raised SNR

Same as Expt. 1 except the stimuli came from the close-talking microphone for the target voice, which raises its signal-to-noise ratio by a significant albeit unknown amount. 60 Ps, none serving in Expt 1.

3.3 E3 Polytextual but not polyphonic:
(a) Homophonic motet: H-5ths, H-Cons, H-Diss
(b) Jingle J-Diss, J-solo: 3-part dissonant, or solo

Intelligibility was so poor in E1 & E2 that E3 tested musical influences using the same lyrics, but only 12 keywords (one set rather than two), and only the Vdiff competitors e.g. keyword *lamb* vs. *limb*. The aim was to assess intelligibility in less challenging musical conditions. Of 5 conditions, 3 manipulated motet *harmonization* to reflect principles of auditory streaming and voice-leading [10, 11, 15, 16]. H-5ths had modal tonality, but a homorhythmic texture, its parts moving in consecutive octaves, fifths and fourths, B and A an octave apart, and T a fifth higher than B. A homophonic Consonant condition, H-Cons, diatonicized the melodies by giving them the closest key to their mode and removing accidentals, then gave them consonant homorhythmic 6-3 chords by transposing T down a 4th and B down a 6th e.g. CGE (from the highest pitch). These chords were consonant to avoid sensory dissonance, and diatonic to avoid harmonic dissonance. H-Diss was the same except dissonant: original melodies were transposed down a tritone to form the T part, and down a minor seventh to form the B part e.g. C F# D#. Melodies were the 4 from E1, plus another 8 composed by the second author in the same style so that each sentence had its own melody, range 5-10 bars.

In J-solo and J-Diss, the texts were set to 12 new alto melodies with a *jazz-inflected diatonicism*, and a 140 bpm 4/4 metre roughly approximating speech rhythms. Pitch range was narrower (rarely more than a fifth), pitch height, syncopation and accents emphasized word boundaries, and cadences and rhythm divided texts into much smaller chunks than the motet conditions. J-Solo was sung by Alto alone. J-Diss stimuli were harmonized as for H-Diss.

Voices were recorded separately, B first. An alto (classically-trained 2nd author) sang A and T parts, hearing the other recorded voice(s) via headphones for tuning and synchronization. A phonetician (first author) judged consistency and quality. The 3 tracks were mixed in Logic Pro, adding a small amount of stereo separation and reverberation for naturalness.

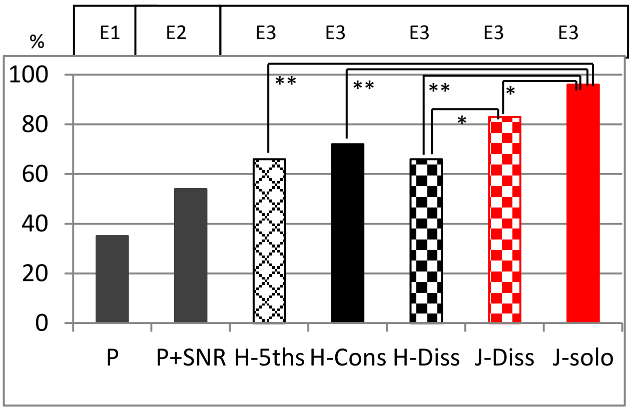
Ps were 11 students aged 18-21, characteristics as for E1. Procedure was as for E1, except the design was fully-crossed repeated measures factorial, and each stimulus was heard 3 times. Stimuli were randomised separately for each P and condition: 3 blocks of 12 = 36 stimuli per condition, 5 conditions (3 homophonic and the 2 jingles) = 180 stimuli total.

4. RESULTS

Phonetic parameters were analysed in full-model logistic regressions for E1 & E2. E3 used ANOVA: 5Condition x 2AC x 2VL x 3VQ repeated measures. Fig. 1 shows percent correct. Columns for E1 & E2 confirm very low intelligibility for Polyphonic motets especially in the standard style: 35% correct. Accuracy rose 19% for E2's favourable P+SNR, but was still only 54%. E3's Homophonic conditions were 66-72% intelligible, despite the same person singing A and T parts, unlike in E1 and E2. Jingles were quite intelligible (polytextual J-Diss 83%, J-solo 96%). In E3, Conditions differed (main effect p<0.001). J-solo was better than all others (p≤ .025), and J-Diss > H-Diss (p=0.037). So the main problem seems to be the motet genre and articulation style rather than singing *per se*, polytextuality (competing words) or the strongly semantically-anomalous texts.

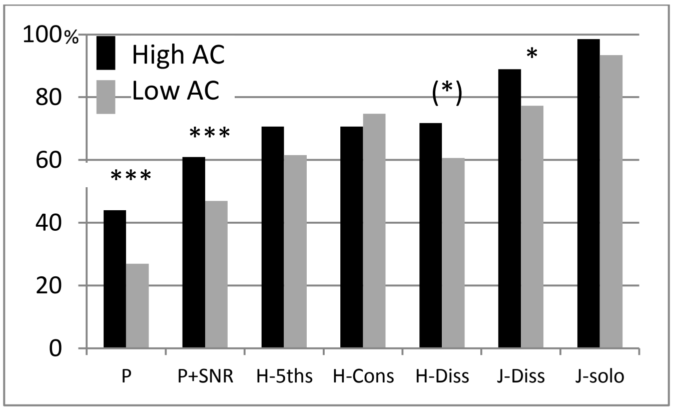
Competitor word affected intelligibility in E1 but not E2, presumably due to E2's enhanced SNR. E1's main effect was significant ($\chi^2(1) = 6.6$, p = 0.037): best intelligibility for Odiff, 39%, Vdiff worst, 30%; Cdiff 36%. Vdiff was predicted to benefit least, for strong masking is expected from different vowels of the same length; Odiff and Cdiff were expected to benefit more, as, respectively, they had identical target and competitor rhymes, or onset and nuclei. But Competitor word also interacted with other factors in E1 & E2, notably AC (see below).

Figure 1: Percent correct responses, non-phonetic conditions, E1-3. Genres: P, H: polyphonic, homophonic motet; J: jingle. * $p \leq 0.05$. ** $p \leq 0.01$. *** $p \leq 0.001$.



High Acoustic Contrast was the most consistently beneficial phonetic parameter (Fig. 2). Gains were significant for E1 and E2 (each p < 0.001) For E3, AC x Condition was p=0.04, with J-Diss p=0.03, H-Diss marginal (p=0.07) and J-solo at ceiling.

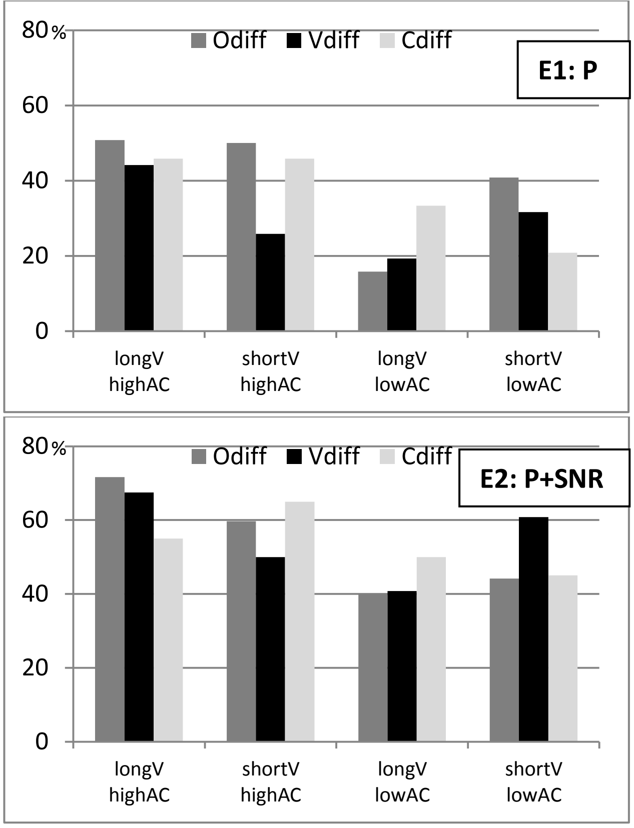
Figure 2: Percent correct responses for High (black) vs Low (grey) Acoustic Contrast. See Fig. 1 for axis details.



This pattern suggests AC may be most influential in more challenging genres, P and P+SNR, and when harmonies are dissonant, but these conclusions are tentative for 3 reasons: results for H-Cons may be artifactual (see Discussion); in E1 & E2, Competitor word interacts with AC, differences being smallest for Vdiff (which is E3's only competitor condition), and both factors form a 3-way interaction with VL.

Vowel Length. The main effect of VL was not significant in the polyphonic E1 and E2, but was significant in E3 (long > short, p =0.021), interacting with Condition (p = 0.023). In pairwise comparisons only J-Diss was significant (p < 0.001), with J-solo marginally so (p = 0.067)—perhaps a ceiling effect.

Figure 3: Competitor word x AC x VL interactions



Interactions of VL with Competitor word and AC suggest that VL is of secondary benefit to intelligibility. Fig. 3 shows 3-way interactions for E1 P: ($\chi^2(2) = 21.2$, $p < 0.001$) and E2 P+SNR ($\chi^2(2) = 5.7$, $p = 0.017$). Short vowels disproportionately benefitted word-spotting when AC was low, especially in difficult Vdiff conditions ($\chi^2(1) = 12.7$, $p < 0.001$); $\chi^2(1) = 11.6$, $p = 0.001$). Conversely, high AC long vowels scored best overall and low AC ones scored poorly. With its higher intelligibility and only Vdiff, E3's AC x VL interaction was not significant (p = 0.122), but to the extent there was a difference, it took the opposite pattern to that of the polyphonic conditions: in E3's homophonic music, short vowels at low AC scored low, 67%, whereas the other 3 conditions were 80%. Finally, that Competitor x VL was significant in E1 (p < 0.001) but not E2 (p = 0.68) further supports the interpretation that VL influences intelligibility in polyphony when blend is good i.e. SNR is poor.

Vowel quality produced many significant effects, but with inconsistent patterns. In E1 & 2, these involved Vdiff in all combinations of VQ with AC and VL, and Odiff in some combinations of VQ with Low AC and Long Vowel, but no clear pattern emerged. In E3, short (high) Front vowels tended to be least intelligible, but its effects were inconsistent in interaction with other factors.

4. GENERAL DISCUSSION

To examine acoustic influences on intelligibility of sung words, this study used monosyllabic real words sequenced to have no sentential meaning. Musical properties seem more powerful than phonetic, though phonetic properties can enhance effects of musical choices. Anomalous text can be intelligible in the right conditions (jingle, limited word choice, no competing text), but polytextual polyphonic motets are largely unintelligible even at good SNR, unlike spoken isolated monosyllables in noise [13].

Good intelligibility arises when words are easy to segment. The best way to achieve this is by clear, text-appropriate rhythm, rests, accents (and good articulation), as in the Jingle style, but high-acoustic-contrast word choice can enhance musical effects, presumably because voiceless obstruents, like the musical factors, introduce f0 discontinuities at word boundaries, but not large pitch jumps, which hinder.

Differences between the polyphonic E1 and E2, and homophonic/solo E3 show that compositional style and texture influenced intelligibility more than timbre and pitch differences between voices. Stream segregation [2] would predict that E1 & E2, with a female target and male competitor, would be easier than E3, where the same person sang both target and competitor, but the reverse was found (even though timing in E3 was so tight that sounds sometimes migrated perceptually between parts e.g. T *cast* and A *mile* sounded like *kyle*). That intelligibility was greater in homophonic than polyphonic conditions may have been due to conjoint part movement during chord progression, which may help to stream melody from lower parts [2, 10]. Huron ([10] & pers. comm. 2013) predicts that *tonal fusion* as in H-5ths facilitates intelligibility when words are the same, but might hinder when they differ, as here. However, that H-5ths and H-Cons did not differ may be due to anomalies in the H-Cons condition. These are as yet unidentified, but are being investigated.

Dissonance is not predicted to hinder [10], and H-Diss and J-Diss combined sensory and harmonic dissonance. Pilot work distinguishing them shows no intelligibility difference, but these and other data suggest dissonance effects may be complex.

Inconsistency between experiments is not readily interpretable. Much may depend on the setting (and melody) and singers' production, but no consistent listener differences were found. In practical terms, this is good news: choose words freely but place them in predictable contexts [8]; and to understand motets, take a copy of the score and words with you.

5. REFERENCES

[1] Benolken, M.S., Swanson, C.E. 1990. The effect of pitch-related changes on the perception of sung vowels. *Journal of the Acoustical Society of America* 87, 1781-1785.

[2] Bregman, A.S. 1990. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

[3] Carlsson, G., Sundberg, J. 1992. Formant frequency tuning in singing. *Journal of Voice* 6, 256-260.

[4] Collister, L.B., Huron, D. 2008. Comparison of word intelligibility in spoken and sung phrases. *Empirical Musicology Review* 3, 109-125.

[5] Fine, P., Ginsborg, J., Barlow, C. 2009. The influence of listeners' singing experience and the number of singers on the understanding of sung text. In: Williamon A., Pretty S., Buck R., editors. The influence of listeners' singing experience and the number of singers on the understanding of sung text. International Symposium on Performance Science: European Association of Conservatoires (AEC).

[6] Ginsborg, J., Fine, P., Barlow, C. 2011. Have we made ourselves clear? Singers and non-singers' perceptions of the intelligibility of sung text. In: Williamon A., Edwards D., Bartel L., editors. Have we made ourselves clear? Singers and non-singers' perceptions of the intelligibility of sung text. International Symposium on Performance Science: European Association of Conservatoires (AEC).

[7] Gregg, J.W., Scherer, R.C. 2006. Vowel intelligibility in classical singing. *Journal of Voice* 20, 198-210.

[8] Heinrich, A., Knight, S., Hawkins, S. under revision. Influences of word predictability and background noise on intelligibility of sung text in live concerts. *Journal of the Acoustical Society of America*.

[9] Hollien, H., Mendes-Schwartz, A.P., Nielsen, K. 2000. Perceptual confusions of high-pitched sung vowels. *Journal of Voice* 14, 287-298.

[10] Huron, D. 2001. Tone and Voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception* 19, 1-64.

[11] Huron, D. 2008. Asynchronous preparation of tonally fused intervals in polyphonic music. *Empirical Musicology Review* 3, 2008.

[12] Johnson, R., Huron, D., Collister, L.B. 2014. Music and lyrics interactions and their influence on recognition of sung words: An investigation of word frequency, rhyme, metric stress, vocal timbre, melisma and repetition priming. *Empirical Musicology Review* 9, 2-20.

[13] Miller, G.A., Heise, G.A., Lichten, W. 1951. The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology* 41, 329-335.

[14] Smith, J., Wolfe, J. 2009. Vowel-pitch matching in Wagner's operas: Implications for intelligibility and ease of singing. *The Journal of the Acoustical Society of America* 125, EL196-EL201.

[15] Wright, J. 2008. Commentary on "Asynchronous preparation of tonally fused intervals in polyphonic music" by David Huron. *Empirical Musicology Review* 3, 69-72.

[16] Wright, J.K., Bregman, A.S. 1987. Auditory stream segregation and the control of dissonance in polyphonic music. *Contemporary Music Review* 2, 63-92.

Supported in part by a Cambridge University Newton Trust Small Grant and the Faculty of Music, Cambridge. We would like to thank Edward Wickham for composing the medieval motet music for E1 (and hence E2), arranging for E1 to be sung, and for singing the bass parts in all experiments. We are also indebted to Yvonne Flory and Stephan Gericke for help with programming and running subjects, and to Jessica Dandy and Gwilym Bowen for singing alto and tenor respectively in E1 and hence E2.