# From Speech to Song: A Response to Johnson, Huron and Collister on the Interaction of Music and Lyrics

EDWARD WICKHAM
*St Catharine's College, University of Cambridge*

ABSTRACT: This commentary addresses some of the methodological difficulties inherent in studying intelligibility in sung text. It provides observations which complement and critique some of the authors' hypotheses and conclusions, and offers a caveat to further 'bottom-up' studies of text intelligibility in music.

Submitted 2013 September 5; accepted 2013 October 24.

KEYWORDS: *intelligibility, singing, lyrics, text setting*

## TESTS FOR SPOKEN AND SUNG WORDS COMPARED

WHETHER it is apocryphal need not concern us; the story has about it an authenticity which is provocative. It tells of a Diva who, frustrated by her audience's focus not on the stage but on the texts printed in their programme booklets, breaks off mid-song and asks of her public which of them can understand the English language. It being an English concert hall, all but a small minority put up their hands. "Then get your heads out of your programmes and listen," she demands imperiously, before continuing with the recital. The singer to whom this anecdote was attached – in the version related to me – was somebody well-known for her crisp articulation and powers of communication; and one has every sympathy for the performer who yearns for eye contact, but manages instead only to project to the often greying or reflective surfaces of the audiences' heads.

Yet one can also imagine being a member of that audience who, while being able clearly to make out each phoneme, each syllable, each word, might still feel the need to understand how all these bits of language fit together into a coherent linguistic unit – a phrase, a sentence, a couplet even. Making phonemes and syllables intelligible is one thing; conveying the meaning of a self-sufficient linguistic unit is something very different. Perhaps unwittingly, the authors suggest a distinction between these two forms of understanding when they use in the title of their paper the word 'recognition' but subsequently adopt the term 'intelligibility'. The latter might be an apt term for the correct processing of language on the level of phonemes, syllables and mono-syllabic words; while the understanding of larger-scale linguistic units requires collaboration between that same small-scale intelligibility and 'recognition', through a multiplicity of contexts, of the linguistic patterns and semantic tropes which enable us to understand a text.

The contexts which enable us to 'recognise' or understand speech – and, crucially, enable us to continue doing this when some of the signal is obscured by extraneous noise – include everything from the thematic (an expectation of the sort of language which might be used in a particular situation), to the physical (the gestures and lip movement of the speaker), and much else besides. The work of Coleman and Hawkins (e.g. Hawkins, 2003; Heinrich & Hawkins, 2009; Coleman, 2003) amongst others, have indicated how we should be considering these contexts on a much more detailed level as well. The quality of a vowel sound alters according to the consonants that precede and follow it: Figures 1a and 1b in the paper under review happen to provide a good example of this. Speak the words 'really' and 'release' to yourself, and you will note the subtle difference in the vowel sound of the first syllable. Not only that, but these vowel sound changes can be detected even when the phonetic differences are not local to the vowel sound in question (Coleman, 2003).

The 'extraneous noise' referred to above may take the form of background babble, machine noise or a sudden alarm; in the case of music, it might be of the more harmonious kind – a piano accompaniment, or voices *a cappella*. It is rare that we have the opportunity to communicate in an entirely noise-free environment; and observations of word intelligibility made by speech scientists working with background noise point to a situation of such bewildering variability, that it encourages an even stronger belief in the power of context, on all levels from the phonetic upwards, as a means of understanding speech (Wang & Bilger, 1973). The resources we bring to bear on a speech signal are multi-modal and include the processing not just of phonetic details, but of patterns of pitch and rhythm in the speech signal to which we are listening. In short, we make use of every type of information we can get.

Nevertheless, it is worth considering to what extent these observations, which refer to speech, can be translated into the study of sung text; and one of the many illuminating aspects about the paper under consideration is its use of a methodology that examines the two side by side. For music more or less obliterates those pitch and rhythmic patterns which make speech so predictable, replacing them by pitches and rhythms which – however pleasing – are more uniform and far less nuanced. In the process, vowels sounds are distorted and consonants often suppressed. A classical singer in the Conservatoire tradition will invariably be taught to prioritise melodic line over textual clarity, and supplant the vowel sounds of normal speech with those most resonant for a particular pitch.

Yet this does not make the situation any easier. The experimental parameters which must be patrolled when putting together tests for intelligibility in such text are numerous; and one must necessarily make choices. Vocal timbre is an obvious one: and the authors have sensibly picked singers from operatic and music theatre backgrounds as representing two different traditions of vocal production. The result – that there is no discernible difference in intelligibility between the two – is reassuring. If one could show that this held for other vocal timbres as well – for instance, falsettists and choral sopranos of the type favoured by early music polyphony choirs – then one might go some way in addressing a widely-held prejudice against the use of vibrato in pre-Classical repertories.

## METHODOLOGY

In keeping with so many speech intelligibility tests, the tests documented here, and in the authors' previous article (Collister & Huron, 2008), entail target words at the end of a phrase, with each target word introduced by the phrase 'I am singing/speaking the word …'. While there are good methodological reasons for doing this, two concerns arise from this approach. The first involves the rhythmic positioning of the test phrase and of the target word itself. I suspect that, in monosyllabic words (as in bisyllabic – see Hypothesis 4), a change in stress pattern will change the articulation and pronunciation of the whole phrase, including the target word. However carefully instructed the singers were for these tests, one cannot be sure that the target word was not articulated differently depending on whether it fell on a weak or strong beat. More significantly, the requirement that target words come at the end of phrases means that there is no sense of how important might be the confounding influence of a succeeding word. It would be hard-hearted in the extreme to criticise such a project on the grounds that the conditions were unrealistic – how else can any such research be pursued except through abstraction? – but in the light of observations to be made below regarding Hypothesis 6, this is a consideration worth bearing in mind.

## HYPOTHESES AND OUTCOMES

That listeners use every type of cue available to make sense of a verbal signal, be it spoken or sung, is securely borne out by these tests. In particular, the results of Hypotheses 2 and 3 demonstrate how greater complexity in the target word (bisyllabic rather than monosyllabic; diphthong rather than monophtong) ensures greater lexical distinctiveness and thus higher intelligibility. In the case of H3, I suspect that the average notes assigned to each syllable in a melisma is of negligible importance by comparison to the number of syllables, though this would need to be tested. H2 is a more interesting case; and, as a singer, choral conductor and sometime singing teacher, I would have made the same assumption as the authors: that sung diphthongs are less intelligible than sung monophthongs. The opposite is the case; a result which reminds us that a truly 'mono' monophthong is rare. The challenge of creating satisfactorily resonant sounds on even mid-range vocal pitches will, in the case of most well-trained singers, result in some kind of vowel distortion. Singing teachers often train their pupils to develop a palette of vowel sounds which is consistent in its tone colour throughout the vocal range, and which suits the vocal timbre of that particular singer, rather than requiring sung vowels to resemble their spoken counterparts. At the same time, the kinds of small-scale distortions of vowel in combination with particular consonants (cited above), can be exaggerated when sung. Take again the example of 'really' and 'release', sung at mid or high range. The inclination to create a diphthong in 'really' is strong, whereas 'release' is more likely to remain a monophthong. An anecdote pertinent to this discussion of creating diphthongs from monophthongal words – in relation to regional accents in particular – was related to me by a singing coach working in Texas. Having lectured his pupil for some time about the importance of controlling one's vowel sounds, the lady asked (and here I attempt a free transliteration of a thick Texan accent), 'Say, mister: what is a dee-ap-tha-wunggg?'

Hypothesis 4 deals with the matching of word stress to musical stress, and the outcome from these tests is predictable though valuable none the less. One additional observation here might be that composers, like poets, frequently exploit the tension between word and musical stress, and the intuitive response of singers might be to emphasise these 'mis-placed' syllables in such a way as to render them more, rather than less intelligible.

The outcome of Hypothesis 6 is the most thought-provoking. Like the authors, I would again have assumed that, when a word is repeated, a format of syllabic followed by melismatic statements would render the target word more intelligible than the reverse format. To the authors' speculative explanation I would add another, concerning the importance of final consonants for intelligibility. A mid-phrase target word is highly likely to be affected by the word succeeding it, whether or not that succeeding word is the same or different and whether or not the singer takes a breath between one word and its successor. In particular, the final consonants of the earlier word will be slightly suppressed, confounded or shortened in order to make way for the succeeding word. This degradation of final consonants is particularly obvious when the succeeding word begins with a consonant. If the succeeding word begins with a vowel, there is a problem of possible elision between consonant and vowel that might be exaggerated depending on the pitch interval between the two syllables. Assuming – as the results of H3 inform us we can – that it is the syllabic rendition of the word rather than the melismatic rendition that provides the clearer clues to what the word is, then listeners are liable to lose more clues from the syllabic rendition of the word appearing first. The obvious objection one might make against this conjecture is that, while one order would favour the intelligibility of the start of the target word, it would disadvantage the end of the word and vice versa. However, with the tests set up as they have been, with each target word introduced by the phrase 'I am saying/singing the word …', I would suggest that the singers would naturally and even unconsciously take special care articulating the start of the target word.

## CONCLUDING REMARKS

The authors' paper does all of us who work with vocal music a great service, whether they be composers or performers, helping to elucidate some issues which have been understood on the instinctive level and offering useful correctives to some misguided intuitions. While in the territory of intelligibility in texted music small parcels of land are being delineated and mapped – I myself am working on a project examining intelligibility in polytextual music[1] – much of this is still *terra incognita*, and we are all at the moment setting out triangulation stations.

But before embarking on ever more elaborate and detailed studies of the intelligibility of sung text, it is important that we get a firm grip on our ambitions and expectations. When I started working on the *Tales from Babel* project with my vocal ensemble, The Clerks, I was horrified to discover how little audiences could hear of the words we so diligently articulate. The second shock was to learn how few people were really concerned how much of the words they understood. What constitutes a 'meaningful' musical experience for them goes way beyond – perhaps even circumvents – the understanding of texts. Different musical traditions, genres and circumstances entail different assumptions about the importance of text intelligibility, and the 'meaningfulness' of lyrics to a listener changes as the listener becomes familiar with a particular song (Thompson & Russo, 2004). But it is instructive to remember that the lack of concern for text intelligibility extends to popular and folk traditions, as is not confined to vocal polyphony in foreign and/or dead languages. The phenomenon of the 'mondegreen' – mis-hearings of popular song lyrics – suggests that listeners are entirely comfortable with nonsensical versions of cherished songs.

This is where the analogy with speech intelligibility breaks down. For all the importance of context and physical gesture, successful speech communication still requires a significant percentage of words to be understood; as the tourist knows all too well as he or she feverishly gesticulates to the foreign interlocutor, actual words matter. This is not the case when words are set to music. Not only does the act of singing, and the vocal techniques involved, obscure the words in multifarious ways; but the music diverts our attention, seduces our intellect (as many a theologian has complained), and perhaps even induces in us an evolutionarily determined instinct to prioritise one kind of message over another.

However sophisticated the tests one might create for intelligibility in sung text, in however realistic a performance environment such tests are conducted, and however well controlled are the parameters within the methodology operates, is it really possible to overcome this basic perceptual lethargy?

## NOTES

[1] A project description, with sample tests and other materials, can be found at www.talesfrombabel.co.uk.

## REFERENCES

Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics*, *31*(3-4), 351-372.

Collister, L., & Huron, D. (2008). Comparison of word intelligibility in spoken and sung phrases. *Empirical Musicology Review*, *3*(3), 109-125.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3-4), 373-405.

Heinrich, A., & Hawkins, S. (2009). Effect of r-resonance information on intelligibility. In *INTERSPEECH* (pp. 804-807).

Thompson, W. F., & Russo, F. A. (2004). The attribution of meaning and emotion to song lyrics. *Polskie Forum Psychologiczne*, *9*, 51-62.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: a study of perceptual features. *The Journal of the Acoustical Society of America*, *54*(5), 1248-1266.