

# A comparative study of spoken and sung voice in performance

Jean Callaghan

Independent Researcher, Sydney, Australia  
jean.callaghan@bigpond.com - <http://users.bigond.net.au/CallaghanSingingVoice/>

Edward McDonald

School of Asian Studies, The University of Auckland, New Zealand  
e.mcdonald@auckland.ac.nz - <http://www.auckland.ac.nz/>

In: K. Maimets-Volt, R. Parncutt, M. Marin & J. Ross (Eds.)  
Proceedings of the third Conference on Interdisciplinary Musicology (CIM07)  
Tallinn, Estonia, 15-19 August 2007, <http://www-gewi.uni-graz.at/cim07/>

**Background in singing voice.** Singing performance involves the expression of musical and linguistic features of a complex musico-verbal text (Callaghan & McDonald 2002) through vocal tone and word articulation. Western classical singing follows the basic maxim of the Italian tradition of vocal pedagogy "One sings as one speaks", and classical singers are trained using the Italian language in order to facilitate vocal resonance and minimal interference of consonants, a feature of that language. This tradition becomes problematic, however, when applied to a language like English which shows a different balance of vocalic and consonantal articulation. Voice science now offers acoustic analysis of aspects of the vocal sound such as pitch, loudness, and timbre (onset, vibrato, and vocal harmonics), clarifying the differences between spoken and sung language, and allowing a more nuanced understanding of how they might be combined in performance.

**Background in linguistics.** Detailed descriptions have been developed within linguistics of the articulation and phonation (intonation) patterns involved in spoken language (Catford, 2001; Ladefoged, 2006). Phonetic studies on spoken language tend to concentrate on articulation rather than phonation, while those on sung language tend to the opposite bias, due to their greater respective elaboration in those areas. The normally assumed "Italianisation" of vocal articulation in classical singing (Nair, 1999), introduces a further bias by ironing out the more complex consonantal combinations and vowel distinctions of spoken English

**Aims.** Although previous research has addressed broad differences between speech and singing (Miller, 1996; Nair, 1999), it has not done so in a specific performance context. The current research aims to clarify the distinctive differences between spoken and sung voice, and problematise the accommodation necessary between them in performance, through a detailed acoustic comparison of spoken and sung versions of the same English text.

**Main contribution.** The poem "Old Sir Faulk" was set to music twice, firstly spoken over instrumental accompaniment (Sitwell & Walton, 1922), and subsequently arranged as a song with piano accompaniment (Walton, 1932). The current study takes as its data the vocal part of each version as performed by the second author in his native Australian accent. The basic rhythmic parameters of both versions are determined by the musical accompaniment, while the articulatory features are held constant by the use of an authentic spoken dialect of English, allowing the more specific contrasts to emerge more clearly.

**Implications.** The performance of a song represents a compromise between the musical and linguistic features of the complex musico-verbal text. A clearer understanding of the nature of both should allow singers to strike a better balance between the two in performance, with positive implications for pedagogy. Text-based analyses such as this study also provide a controlled context for musicologists and linguists / phoneticians to explore the extent of mutual influence between spoken and sung vocal features within a particular style and across different musical styles.

Because language and music are two semiotic systems which use sound as their medium of expression, numbers of scholars have attempted to build models identifying the links between them; for example, Burrows (1990), van Leeuwen (1999), and various studies in Wallin, Merker & Brown (2000). The fundamental sound features common to both language and music are variation in duration

(rhythm) and variation in pitch (melody). The similarities and differences in their employment of these sound features are thrown into sharp relief in song, which draws on both systems. Singing performance is concerned with conveying both musical and verbal meanings, expressed in musical and linguistic features of the text, and realised in

performance through vocal tone and word articulation.

Both speech and music are heard in terms of the four perceptual categories of pitch, loudness, duration and quality. But these are subjective categories which cannot be equated exactly with the physiological or physical categories of vocal-fold vibration/fundamental frequency; breath effort/intensity; physical duration; or shape of vocal cavities/spectral structure. While the quality of the singing voice depends essentially on its source, vocal-fold vibration, it also depends on how the spectrum of sound thus produced is filtered by the vocal tract. In singing, articulatory adjustments (for both timbre and word articulation) need to be achieved without compromising the voice source. Voice source and vocal tract filter are interdependent, just as subglottal pressure and laryngeal adjustment are interdependent (Callaghan, 2000). Vocal resonance and word articulation are both reliant on the movement of the articulators: The jaw, soft palate, lips and tongue. In speech, those same articulators are used to produce the vowels, as well as the different types of consonants, which to a greater or lesser extent interrupt the vocal tone. When combined in singing, word articulation may require modification in order to maintain the vocal tone appropriate to the musical style.

In discussing the adage from the historic Italian school, "*Si canta come si parla*" (one sings as one speaks), Miller (1996) points out that singing is not simply sustained speech spun out over wide-ranging pitch fluctuations. It differs in terms of vowel definition, phonetic duration, and intensity. Major differences between speech and singing are: The need in singing to maintain vowel integrity over a wide pitch range; the length of time occupied by phonemes; the avoidance or minimisation of transitions between phonemes in singing; the increase in intensity in singing.

In a more recent publication, Nair identifies in greater detail seven ways in which singing is unlike speech (2007, pp. 6-9):

1. Vowels, and some consonant phonemes, are sustained for longer than those in speech.
2. All phonemes, especially vowels, are performed more richly (more resonantly) than those in speech.
3. Individual sung phonemes are generally purer and more consistent in song than those in speech.
4. All singing phonemes are joined with the surrounding phonemes in far more precise ways than in speech.
5. Singing sounds are usually performed at greater volume than those of normal speech.
6. Song is performed with a pitch range that far exceeds the norms of speech.
7. Sung phonemes are executed with a rhythmic accuracy determined by the composer's musical notation.

Even in a piece that attempts to follow the speech rhythms of the text, these differences can usually be observed. There is, however, a need to be more specific about what features are common to speech and singing, which differ, and how they differ.

### **Aim**

In order to identify features common to speaking and singing and those that differ between the two, the current research analysed a spoken and sung version of the same linguistic text. The aim was, given the same configuration of phonemes in the same rhythm, to identify how the vocal realisation differed in respect to resonance, vowel-consonant transitions, intensity, pitch range and rhythmic accuracy.

### **Nature of the Analysed Text**

The basic verbal text is a poem by Dame Edith Sitwell, set by William Walton in two versions. The original spoken version was first performed in 1922 as one movement of a multi-movement performance piece for "reciter" and six instruments, called *Façade—An Entertainment*, where the poem was read out in rhythm over the musical accompaniment. The "Old Sir Faulk" text was used for "Fox-trot", the twentieth of 21 short movements. In the subsequent sung version with piano accompaniment (1932), titled "Old Sir Faulk", Walton retained substantially the

same rhythms, and fitted the words to a melody line based on the original instrumental accompaniment.

The concept of *Façade* demonstrates at least three contemporary influences: Cabaret, the traditions of melodrama and *Sprechstimme* / *Sprechgesang*, and the musical possibilities inherent in the English language. "To the cabaretists of the turn-of-the-century, art could best free itself from the tyranny of commitment by vigorously asserting its right to purpose-lessness, its freedom not to have meaning in any conventionally representational sense, its legitimacy as play" (Segel, 1987, p. xxiii). The poems set to music in *Façade* were playful explorations of the sound possibilities of the English language, in contrast to what Sitwell characterised as "the rhythmic flaccidity, the verbal deadness, the dead and expected patterns" (Sitwell 1957: xvi) of some of the poetry of the time. As with the tradition of melodrama and *Sprechstimme/Sprechgesang*, the voice part was declaimed against an instrumental accompaniment. In setting the text as a song, Walton precisely notated all pitches, as well as the rhythm, including accents and articulation. This precise word-setting probably derives partly from Sitwell's highly self-conscious use of sound in the original text, and partly from the practice of the sixteenth-century English madrigalists, who were a strong influence on composers in London at this time (Collaer, 1961, p. 382).

Sitwell characterised the texts in *Façade* as "abstract poems"— as patterns in sound, experimenting with the effect of using rhymes, assonances and dissonances on rhythm and speed. She was also interested in variations in texture achieved by what she called the "thickness" and "thinness" of changing con-sonants, e.g. from a stop to its corresponding fricative ("apiaries" to "aviaries"). This sound play often necessitates distorted syntactic patterns with many run-on lines. Sitwell also specified that the poems be read in a monotone. In the first performance, the poems were spoken through a megaphone from behind a curtain, in order to be heard above the instrumental sound and also to deprive the work of any personal quality. It must therefore be noted that these texts are highly mannered, even "artificial",

as an example of spoken language, with implications that will be noted below.

## Method

The current research was based on a comparative instrumental analysis of separate recordings of the spoken and sung text, addressing the following features:

- duration of phonemes
- formant structure—resonance
- formant structure—consistency of vowels
- transitions between phonemes
- intensity
- pitch range
- rhythmic accuracy

Both performances were by the second author, using his normal educated variety of Australian English and his trained baritone voice. Received Pronunciation [Standard Southern British English] is more often used in singing classical music with an English text, but given the origins of the song in more popular forms, and to facilitate comparison between the spoken and sung form, the performer's own idiom was judged appropriate. In the spoken performance analysed here, the decision was taken not to follow Sitwell's directive, but to read the text in an intonation contour. This was done for several reasons: Firstly, a consistent monotone is very difficult to maintain and also affects intelligibility, in that intonation is one of the means of grouping together grammatically-linked segments of the text. It was also felt that, for the purposes of comparison, having an intonation contour would facilitate the contrast between the two versions.

The composer specified the same rhythmic patterns, and roughly the same tempo, for the two versions. Both the spoken and the sung text were recorded phrase by phrase, at the same tempo, in Sing&See™. Sing&See™ is a commercial software package able to show pitch contour, formant structure, and intensity of vocal sound. Each phrase extract (spoken and sung) was saved as a separate WAV file to enable comparison.

The spectrogram was used for a comparative analysis of: Duration of phonemes; formant structure of phonemes (resonance); formant structure of phonemes (consistency); transitions between phonemes; and rhythmic accuracy. The musical staff display and pitch trace were used to analyse the pitch range. The level meter display was used to compare the intensity of spoken and sung versions.

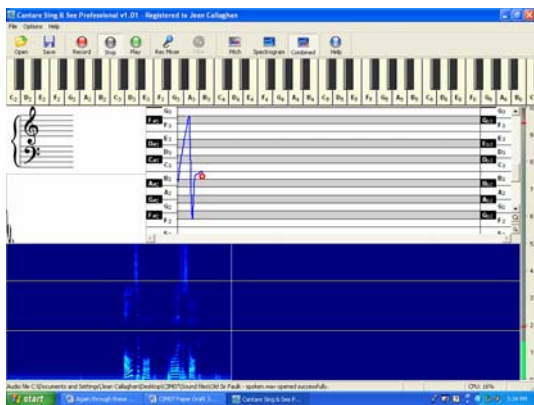
## Results

### Duration of phonemes

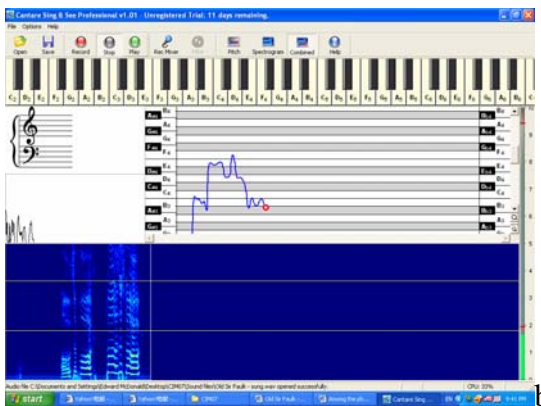
Despite the tempo being the same in both spoken and sung versions, the sung extracts were overall slightly longer than their spoken equivalents, mainly due to the prolongation of final notes of phrases. Within each phrase, in the spoken version the phonemes tended to be shorter and the gaps between them longer. As expected, in the sung version the vowel phonemes were much longer in relation to the consonants.

For example, the spoken and sung versions of the opening phrase, "Old Sir Faulk, tall as a stalk" occupy the same duration, but in the spoken version the consonants take longer than in the sung; in the sung, the vowels are longer (see Figure 1).

Formant structure of phonemes—resonance  
 The sung versions were characterised by more resonant vowels, in that the vowel was clearly defined by the relationship between first and second formant, and that acoustic was sustained for the phoneme. The singer's formant was present throughout, as was consistent vibrato. Figure 2 below shows the spoken and sung versions of the phrase "Seeing the world as a bare egg, Laid by the feathered air", with clearer definition of the vowels in the sung version, consistent vibrato and the presence of the singer's formant.

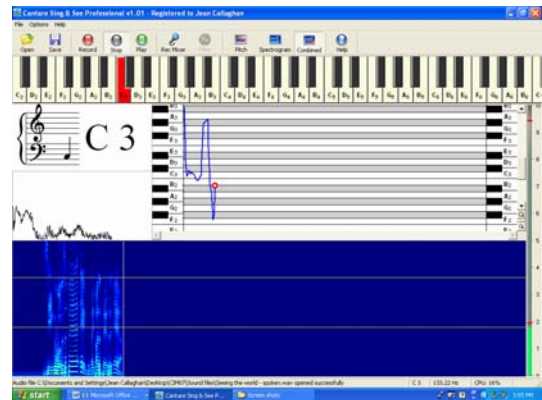


a) Spoken

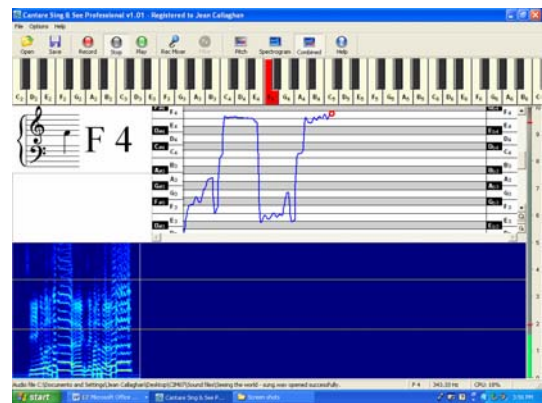


b) Sung

**Figure 1.** "Old Sir Faulk, tall as a stalk".  
 a) Spoken; b) Sung.



a) Spoken



b) Sung

**Figure 2.** "Seeing the world as a bare egg, Laid by the feathered air."  
 a) Spoken; b) Sung.

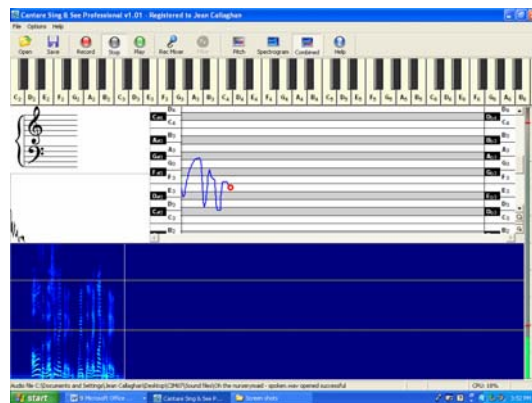
## Formant structure of phonemes—consistency

Vowels were more consistent in the sung than in the spoken version, both during the duration of the phoneme and in comparison with the same phoneme elsewhere in the extract. An exception to this was where vowel modification occurred at high pitch. For example Figure 2 above shows consistent vowels in “bare” and “air” in the sung version, but as these words occur on an F4, high in the baritone voice, they are slightly (but consistently) modified.

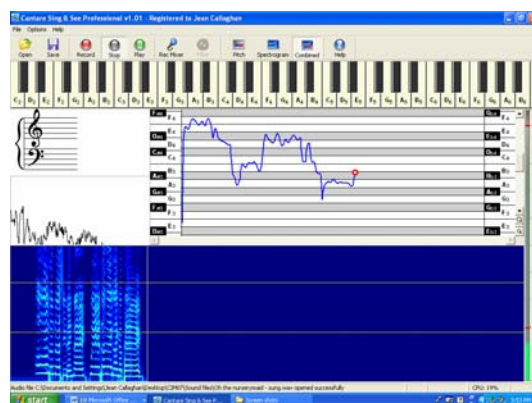
## Transitions between phonemes

Certain transitions were much quicker in the sung version than in the spoken; for example, transitions from consonant to vowel, and from silence to the onset of voice. However, in the sung version transitions from vowel to vowel via a voiced continuant were continuous and maintained on pitch, in contrast to the spoken version where the pitch changed on the transition to the continuant. In the spoken version, pitch also fell on voiced stops. Final consonants were much clearer (shorter) in the sung version. These features can be seen in Figure 3, which displays the spoken and sung form of the phrase “Oh, the nursery-maid Meg, with a leg like a peg”. In the spoken form, the transitions from consonant to vowels in “leg”, and “Meg” are quicker than in the sung form; the pitch drops on the voiced stop /g/ in both words. In the sung form /l/ and /m/ are sung on the pitch of the following vowel, and the final voiced consonants are shorter and clearer.

For diphthongs, the transition from the main vowel to off-glide was quicker in the spoken version, and, depending on its position in the phrase, changed pitch. In the spoken version, pitch on the off-glide tended to descend at the end of a phrase, whereas in the sung version the pitch was sustained throughout head vowel and off-glide. This can be seen in the diphthong /aɪ/ in “like” in Figure 3.



a) Spoken



b) Sung

**Figure 3.** “Oh the nursery-maid Meg, with a leg like a peg.”

a) Spoken; b) Sung.

## Intensity

Because the spoken version was a performance text declaimed over an instrumental backing, intensity was higher than would be expected in conversational speech (showing 0-8 on the level meter display, with an average of 7.5). However, the sung version was higher still (0-10, with an average of 9.5).

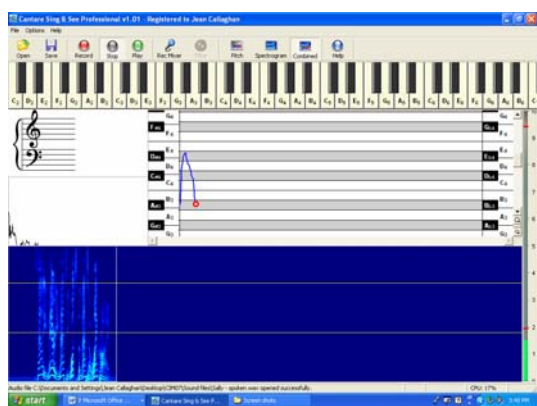
## Pitch range

Miller (1996) suggests that most speaking ranges encompass considerably less than half the fully developed singing range. In the case of this spoken performance version, the pitch range was wider than would be expected in conversational speech. The average pitch range of the spoken version was just over an octave (from A#2 to B3). One exceptional example—the highest in the piece—went up to E4 in the spoken version, because of the declamatory nature of the

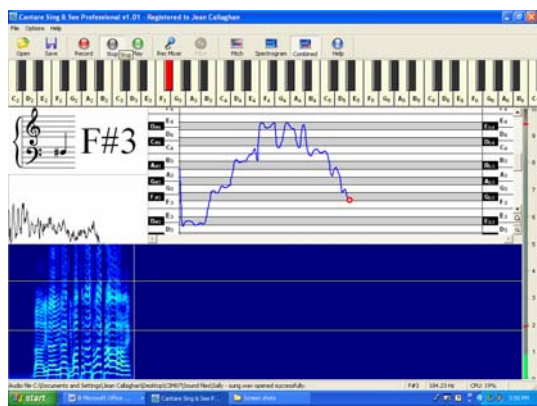
phrase. In the sung version the specific pitches were, of course, dictated by the notated melody line. The overall range was C#3 to G4—a twelfth.

**Rhythmic accuracy**

The concept of rhythmic accuracy was seen to be quite a complex one depending on the types of phoneme segments, particularly consonants, and the transitions and gaps between them: In some cases the spoken version was heard as more accurate; in others the sung. Even in cases where the duration of phonemes was the same in spoken and sung versions, where the rhythm was marked by rests, the fact that the

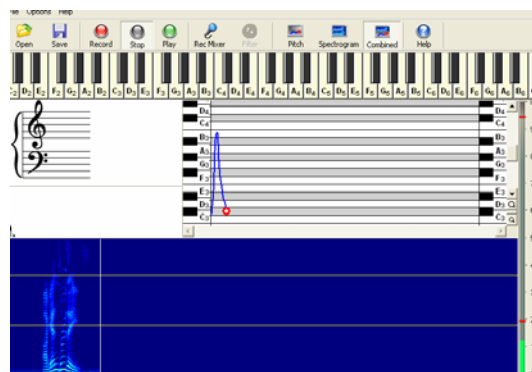


a) Spoken

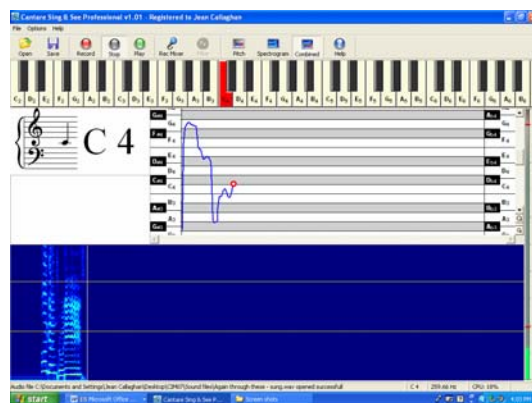


a) Sung

**Figure 4.** "Sally, Mary, Mattie, what's the matter, why cry?"  
a) Spoken; b) Sung



a) Spoken



b) Sung

**Figure 5.** "Again through these."  
a) Spoken; b) Sung

sung version had clearer final consonants and more precise onsets meant that rests were more clearly defined. This made the rhythm more precise. For example, in "Sally, Mary, Mattie, what's the matter, why cry?" (Figure 4), while the overall duration is nearly the same, the rhythm is more accurate in the sung version because of the longer vowels and quicker transitions.

However, in phrases where there are voiced consonants, transitions between vowel and consonant are less clear in the sung version because the voicing continues throughout. In those cases the rhythm of the spoken version is clearer. For example, in the spoken version of the last phrase "Again through these" (Figure 5), the rhythm is more accurate because of the more distinct transitions. In the sung version, the voiced consonants sustained on pitch tend to blur the rhythm.

**Discussion**

As a handy way of summing up the results, we can return to Nair's (2007, pp. 6-9) list of differences between singing and speech quoted above, and in the light of the above analysis, flesh out or modify his claims. The following statements are of course made in relation to the "operatic" or "art song" style of singing and the "declamatory" style of speech adopted here.

1. Voiced segments generally, including vowels, and voiced consonants such as the continuant /l/ and stops /d/ and /n/, are sustained for longer than those in speech, with such consonants typically taking the pitch of the preceding or following vowel.
2. All phonemes, especially vowels, are performed more resonantly than those in speech, with vowels more clearly defined by the relationship between the first two formants, characterised by greater resonance throughout the formant range, including the presence of singer's formant, and by consistent vibrato.
3. Nair's notion of individual sung phonemes as generally "purer" and "more consistent" in song than in speech, needs to be broken down. From the current analysis, this stricture seems to apply mainly to vowels, which as noted above in the light of the formant analysis, are both sustained more consistently within each instance and rendered more consistently between different instances of the same vowel. The main exception to this is at particularly high or low registers in the vocal range, where the vowel quality may need to be modified in order to maintain overall vocal resonance. (In his discussion of speaking vowels versus singing vowels, Titze (1995) suggests three other requirements that may dictate vowel modification in singing: The need for a wider dynamic range; the need to maintain a balance in loudness across phonemes; and a desire to achieve aesthetically interesting vocal qualities.) The notion of consonant phonemes as "purer" would seem to relate largely to the nature of transition, which will be dealt with under Nair's next point.
4. Nair's claim that sung phonemes are joined with surrounding phonemes in ways that are "far more precise" than in speech, relates to the nature of transitions between sounds in speech, which can be characterised by two features: Greater duration of transitions, and greater variability of pitch. As noted above, in the sung version, the transitions between silence and the onset of voice, or between consonant and vowel, are far quicker than in speech, and take place as much as possible on the same pitch. Exceptions to this can occur for expressive effect, such as the quick "run-up" to the main pitch on "Oh" in the sung version of "Oh the nurserymaid Meg" in Figure 3 above. In comparison, transitions in speech tend to take longer, with the transition to and from consonant closure audibly longer and with more "noise" than in the sung version. There is also, depending on the placement in the phrase and therefore the overall intonation contour, much greater variability in pitch in spoken transitions, particularly between two voice segments, for example the main vowel and off-glide of a diphthong, or a vowel and following voiced continuant or stop consonant, where in both cases the pitch tends to fall on the second segment.
5. Nair's claim that singing sounds are usually performed at greater volume than those of normal speech is largely borne out here, except that given the performative nature of the spoken version and the necessity to be heard over an instrumental accompaniment, the difference in volume between the two is not as great as it would be for normal conversational speech.
6. Again Nair's claim that song is performed with a pitch range that far exceeds the norms of speech needs to

be modified in the present case, where the declamatory style of speech calls for an exaggerated pitch range. In addition, the variability of pitch is probably greater in this instance than in ordinary speech: The highly patterned nature of the sound texture, including alliteration, assonance, and both end-rhyme and internal rhyme, as well as the "distorted" syntax and frequent run-on lines, often seem to necessitate a type of "sing-song" intonation.

7. Finally, while Nair's claim that "sung phonemes are executed with a rhythmic accuracy determined by the composer's musical notation" is unexceptionable as phrased, the notion of "rhythmic accuracy" needs to be characterised with more precision. As we saw in the analysis above, the perception of rhythmic accuracy, or perhaps better "rhythmic clarity", seems to be dependent in varying measure on the audibility of the rhythmic beats and on the distinctness of the transitions between beats. Thus, in a case where individual sung beats are longer and have more precise transitions both to consonants and to rests, the perception of rhythmic clarity will be greater. Conversely, where a sung phrase is characterised by greater consistency of articulation, for example a predominance of voiced segments, the transitions between the beats will be less prominent, and therefore the perception of rhythmic clarity will be less, notwithstanding the overall fidelity to the notated rhythm.

### Implications and suggestions for further research

The analysis undertaken in this research was based on a deliberately restricted data set. Both spoken and sung text were performed in largely the same set rhythmical pattern, and with a pitch range and intensity necessary for audibility against a competing noise source (in this case different kinds of instrumental

accompaniment). The fact that despite these commonalities, there were still significant and systematic differences between the two versions, suggests that in the comparison of spoken and sung language, there is the need on the one hand to pay closer attention to the particular variety of each, and that, on the other, the needs of *performed* language, whether spoken or sung, deserve more detailed analysis.

Blanket statements about differences between "spoken" and "sung" language thus need to be broken down into more precise claims about specific vocal styles in relation to particular performative and generic constraints. It would be very informative, for example, to carry out detailed comparison of a series of different vocal styles, stretching from recited speech, to declaimed speech with specific indication for rhythm but not pitch (as analysed here), to *Sprechstimme/Sprechgesang* where rhythm is precisely and pitch approximately indicated, right through to fully sung style (as also analysed here). Equally enlightening results would no doubt emerge from analysing a similar range of vocal styles in a different performance tradition, such as that stretching from rap at the spoken end to rock singing at the sung end. Detailed comparisons, such as carried out in the present study, not only allow researchers to more precisely characterise the range of variation between spoken and sung styles, but should also provide performers and teachers with specific data for feedback and modelling.

**Acknowledgements.** Thanks to Dr William Thorpe, Director of Cantovation Pty Ltd, for his assistance in the use of Sing&See™.

### References

- Burrows, D. (1990). *Sound, speech, and music*. Amherst, MA: The University of Massachusetts Press.
- Callaghan, J. (2000). *Singing and voice science*. San Diego, CA: Singular Publishing Group/Thomson Learning.
- Callaghan, J. & McDonald, E. (2002). Expression, content and meaning in language and music: An integrated semiotic analysis. In P. McKeivitt, S. Ó Nualláin, & C. Mulvihill,

*Language, vision and music* (205-220).  
Amsterdam: Benjamins.

- Catford, J.C. (2001). *A practical introduction to phonetics*. (2<sup>nd</sup> ed.). Oxford: Oxford University Press.
- Collaer, P. (1961). *A history of modern music* (Trans. Sally Abeles). New York: Grosset & Dunlap.
- Ladefoged, P. (2006). *A course in phonetics* (5<sup>th</sup> ed.) Boston: Thomson Wadsworth.
- Miller, R. (1996). *Si canta come si parla?* (47-50); How singing is *not* like speaking (50-52). In *On the art of singing*. New York: Oxford University Press.
- Nair, G. (1999). *Voice—tradition and technology. A state-of-the-art studio*. San Diego: Singular Publishing Group.
- Nair, G. (2007). *The craft of singing*. San Diego, CA: Plural Publishing.
- Segel, H.B. (1987). *Turn-of-the-century cabaret*. New York: Columbia University Press.
- Sitwell, E. (1957). *Collected Poems*. London: Macmillan.
- Sitwell, E. & Walton, W. (1922). *Façade. An entertainment with poems by Edith Sitwell and music by William Walton*. London: Faber.
- Titze, I. (1995). Voice research: Speaking vowels versus singing vowels, *Journal of Singing*. 52(1), 41-42.
- van Leeuwen, T. (1999). *Speech, music, sound*. London: Macmillan.
- Wallin, N.L., Merker, B. & Brown, S. (2000). *The origins of music*. Cambridge, MA: MIT Press.
- Walton, W. (1932). Three songs. Poems by Edith Sitwell; music by William Walton. Oxford: Oxford University Press.