“Softened” Voice Quality in Poetry Reading: An Acoustic Study

*** Author accepted manuscript ***

Full reference:


Sound files for this article:
http://www.tau.ac.il/~tsurxx/SoftVoiceSoundFiles/SoftVoiceSoundFiles.html
“Softened” Voice Quality in Poetry Reading: An Acoustic Study

This study explores the acoustic properties of a certain “softened” voice quality adopted by some professional poetry readers for emotive expression. We show that several acoustic features may serve as cues for softened voice, including: breathy voice, low intensity, and, possibly, reduced tempo and restrained dynamics of intensity and pitch contours. In addition, we show that different speakers may achieve the effect of voice softening in different ways. Our results demonstrate the contribution of the vocal dimension to the experience of poetry. Although this is essentially a case study of the subjective perceptions of one individual, the results are consistent, and we show their potential to be generalized across listeners.

Keywords: Voice quality, Poetry reading, Soft voice, Breathy voice, Spectral emphasis, Emotive speech
Stating the problem

In course of our instrumental study of the rhythmical performance of poetry we noticed that some performers sometimes had recourse to a certain voice quality, deviating from that of the context, for emotive expression. This quality could be roughly described as “softened”, “tender”. Let us provide, first, an ostensive definition of this quality. Listen to an example from the reading of “Kubla Khan” (Coleridge, “Kubla Khan”) by Michael Sheen (Coleridge, “Kubla Khan”).

And all should cry, Beware! Beware!
His flashing eyes, his floating hair

In this example, “and all should cry” displays the standard voice quality, from which the voice quality of “beware” deviates in the softened direction. This quality can be taken to express mental states, such as a reflective state of mind or wonderment, and emotions, such as affection and longing. In this paper, we will explore the nature of this voice quality, its phonetic correlates and its possible aesthetic significance.

Sound and aesthetic qualities

The present project is a pilot instrumental study investigating the relationship between measurable physical aspects of the speech signal and some voice qualities that typically count toward certain emotional qualities of a performance. In a previously-published paper, Gafni and Tsur report a study of naïve listener responses to softened voice quality in two readings of “Kubla Khan”. In addition to such responses to softened voice qualities as “calm, gentle, yearning, affectionate, quiet, and slow”, participants provide, to the same items, also such responses as “magical, mysterious, awed, dreamy, and spooky”. Part of our explanation was as follows.

According to our conception of aesthetic qualities, there is a kind of structural resemblance between soft substances, the “softened” voice quality, and such emotional qualities as magical, mysterious, and dreamy. First, soft substances are more yielding and more “responsive” than solid objects; they are more easily “impressed”. Second, as we will demonstrate in this paper, “softened” voice quality, is typically correlated with “incomplete phonation”: whisper, aspiration1 and breathy voice. These do not display the

---

1 Aspiration is the excessive puff of air produced, for example, with the pronunciation of p in pill as opposed to spill (this can be felt by putting one’s hand in front of the mouth while uttering these words).
full sensory presence of solid ordinary phonation. Third, ordinary consciousness consists
in constructing a stable world from a stream of fluid percepts, where certain things, at
least, are firmly fixed. Altered states of consciousness related to magical, mysterious, and
dreamy, by contrast, are “softer,” more “yielding,” and less stable. The various kinds of
incomplete phonation may suggest vague, unspecified perceptions, or, perhaps, a world
in which objects have an incomplete existence, hence the “mysterious” or “dreamy”
quality. We may expect that the vocal manipulations explored in the present paper would
not generate some specific emotions, but rather some generalized emotional atmosphere
that may be individuated by the semantic component as well as other phonetic elements
as specific emotions and attitudes.

**Speech sounds acoustics**

In this section we are going to present a theoretical framework, within which we will
explain what we mean by the afore-mentioned “softened voice quality”, and how one can
account for it.

Speech sounds can be described in terms of their acoustic properties, including:
*intensity* (or its perceptual correlate, *loudness*), *duration*, *fundamental frequency*, and
*spectrum* (or *overtone structure*). The latter is highly relevant for our discussion on voice
quality. The generation of speech sounds typically involves vibrations of the vocal cords.
The *fundamental frequency*, which measures the frequency of these vibrations, is the main
contributor to the perception of *pitch* (e.g., high-pitched voice results from vocal cord
vibrations at high frequency). In addition to the fundamental frequency, each sound has
*overtones*. Within overtones one may distinguish *harmonics* and *formants*.

Harmonics are whole number multiples of the fundamental frequency. Vibration
at the fundamental frequency generates the first harmonic (fundamental tone), and
vibration at twice the fundamental frequency generates the second harmonic. The
formants are additional overtones resulting from the interaction of the sound with the
vocal tract (i.e., the tongue and the walls of the oral and nasal cavities). The first formant
has the lowest frequency marked $F_1$. Higher formant frequencies are marked $F_2$, $F_3$, etc.

The fundamental tone and its overtones are bundled together into a complex
unitary sound. The frequencies and relative strength of the various components determine
the identity of the sound. For example, vowel quality (e.g., /a/ vs. /i/) is determined mainly
by the values of the first two formant frequencies. These values, in turn, are determined
primarily by the position of the tongue (e.g., lowered or raised) during speech.
The relative strength (or audibility) of the various overtones is associated with voice quality (e.g., clear vs. breathy voice). This is similar to the notion of timbre or tone-color of musical notes. Any note on the scale (e.g., C or Do) has the same overtone structure (i.e., higher harmonics whose frequencies are integer multiples of the frequency of the given note), regardless of the instrument. Nevertheless, when we listen to the same note played on a violin or a flute, we can distinguish between the instruments from what we hear, without seeing them. This is due to difference of timbre between the notes, which is related, among other things, to different relations between the overtones of the note on each instrument.

The spectrum of sound is a “wave-like” plot, which displays the frequencies and amplitudes of the various components of the sound. Figure 1 shows the spectrum of the vowel /ʌ/ represented by the letter u in the word sunless. The component associated with the fundamental frequency (the first harmonic) is labelled as H1. The plot indicates that the speaker uttered the vowel with a fundamental frequency of 113 Hz, typical of adult male voice. The plot also marks the second harmonic of the voice source (H2), which is approximately twice the frequency of first harmonic. Other peaks in the spectral plot represent additional harmonics and formants. The relative heights (amplitudes) of the spectral peaks indicate their relative salience. In this example, the spectrum is dominated by the first harmonic (which has the largest amplitude). The spectrum also shows a property known as spectral slope (or, spectral tilt), which is the general decrease in peak amplitude with the increase of frequency. This feature of the spectrum is considered to be a universal property of natural sounds (Fry).
Figure 1. A typical spectrum of breathy voice. Shown here: the vowel /ʌ/ in sunless (/sʌnləs/). The first harmonic (H₁) is stronger than the second harmonic (H₂). Also note the point marked by the red crosshair (at 2738 Hz) which has a low amplitude. Positive value of the amplitude difference H₁ – H₂ and low amplitude at high frequencies are typical predictors of unaspirated breathy vowels.

One additional attribute of speech sounds that is relevant to our discussion concerns the periodicity of sound waves. Tsur (“Size-Sound Symbolism Revisited”) argues that periodic consonants (e.g., [m], [n]) are perceived as soft, mellow, and appeasing, whereas aperiodic continuants (e.g., [s], [z]) as harsh, strident, turbulent, and the like. Tsur explained this as follows: “Periodic sounds have been described (May and Repp, 1982: 145) as “the recurrence of signal portions with similar structure”, whereas aperiodic stimuli as having “randomly changing waveform”, that “may have more idiosyncratic features to be remembered”. The recurring signal portions with similar structures may arouse in the perceiver a relatively relaxed kind of attentiveness (there will be no surprises, one may expect the same waveform to recur). Thus, periodic sounds are experienced as smoothly flowing. The randomly changing waveforms of aperiodic sounds, with their “idiosyncratic features”, are experienced as disorder, as a disruption of the “relaxed kind of attentiveness” (Tsur, 1992: 44).” (914).

In this respect, the vocal resources of human language are rather limited. The only way for poets to foreground tonal or harsh sounds in their texts is to increase the relative
occurrence of periodic or aperiodic speech sounds. But in vocal performance the reciter may have recourse to, e.g., breathy or creaky voice. Breathy voice is a phonation in which the vocal cords vibrate, as they do in normal (modal) voicing, but are adjusted so that a larger volume of air escapes, producing a sighing sound (Trask, 1996). Such volume of air is perceived as incomplete phonation. In creaky voice, the vocal folds are compressed rather tightly, and vibrate irregularly at 20–50 pulses per second, about two octaves below the frequency of normal voicing, and the airflow through the glottis is very slow (Trask, 1996).

Both breathy and creaky voice consist in irregular pulses. In breathy voice, however, owing to incomplete phonation, irregularity is barely discernible, if at all (listen to an example from Marasek (1997)). By contrast, in creaky voice, owing to complete phonation, the irregular pulses are very much in evidence, giving the creaky voice a harsh quality (listen to an example from Marasek (1997)), compared to modal voice (listen to an example from Marasek (1997)). As we will see in the next section, the notions of breathy voice and the spectrum of vowels are related to each other and to the ‘softened’ voice quality we study.

An acoustic model of ‘softened’ voice

The concept of ‘softened voice’ refers to a perceived, subjective psycho-acoustic effect. As such, it cannot be given a clear, “scientific” definition. Instead we provide only an ostensive definition of the term; that is, by pointing out examples that generate the same subjective perceptual effect. The choice of the term ‘softened voice’ suggests an intuitively, close link to the notion of ‘soft voice’. However, while ‘soft voice’ normally describes some one-dimensional acoustic aspect of speech voice, ‘softened voice’ refers to the psychological effect generated by the speech voice through various manipulations, such as the use of ‘soft voice’ and others. In the remaining of this section, we will follow our intuitions and construct an acoustic model that captures what we feel to be main cues for the perception of ‘softened voice’.

The term ‘soft voice’ is used in the literature, most commonly, as synonymous with ‘quiet voice’ (Holmberg et al.; Sulter et al.). In other words, we expect softened voice to have lower intensity than a non-softened voice (neutral or other). We also expect changes in the intensity contour of softened voice to be more gradual and bounded in a narrower range compared to non-softened voice.
Some studies associate ‘soft’ voice with breathy voice. Laver notes that “The main phonatory characteristic of breathy voice, incomplete closure of the vocal folds, is consistent with Chiba and Kajiyama’s comment … that ‘in “soft voice” the cartilaginous glottis is kept slightly open’ (20-1).” (146). Another direct link between breathiness and softness appears in Freese and Maynard, who state that: “the use of breathy voice contributes to the hearably “soft” or “soothing” character of many bad news deliveries and responses” (202).

The literature proposes various acoustic parameters as possible cues to breathy voice, including high frequency aspiration noise, low degree of periodicity, strong first harmonic, weak higher overtones, and steep spectral slope (Blankenship, The Time Course of Breathiness and Laryngealization in Vowels; Wayland and Jongman; Klatt and Klatt; Laver; Hillenbrand et al.; Hillenbrand and Houde; Keating and Esposito; Blankenship, “The Timing of Nonmodal Phonation in Vowels”). The characterization of breathy voice as having a strong first harmonic and weak higher overtones fits well with the intuition that has guided our study from the beginning. One of us had the good luck to attend, at a private home, a rehearsal of Bach cantatas produced by a harpsichord, a flute, and a countertenor. The singer happened to perform exactly the same kinds of voice-softening we have noticed in Sheen’s reading. In conversation after the rehearsal he seemed to be very much aware of this voice quality. When asked how he produced it, he answered: “I suppose I am switching off some of the overtones”. In light of the foregoing distinctions we assumed that rather than switching off overtones, he was overemphasizing some overtones at the expense of others. Thus, a major part of our instrumental investigation relies on measuring the properties of vowel overtones (harmonics and formants).

Next, our model predicts that phrases uttered in softened voice are expected to have a restrained (or prolonged) speech tempo (Fónagy and Magdics; Freese and Maynard). It is not clear whether prolongation by itself cues perception of softness. It is possible that prolongation emphasizes other acoustic cues for softness, thus enhancing the perception of softened voice.

Finally, we found occasional references to pitch level and intonation in previous accounts of soft voice and breathy voice. Freese and Maynard report a prototypical example of an empathetic bad news delivery: “After this hesitation, Leslie’s prosody changes: Her pace slows, her voice gets softer… and more breathy… and her pitch lowers.” (205-6). By contrast, Fónagy and Magdics observed that tenderness, and maybe
also longing, are produced in higher pitch level than neutral voice (in our model, both
emotions should exhibit softened characteristics). The above conflicting accounts suggest
that the role of pitch level in the perception of softened voice is not entirely clear.²

As for the intonation, or pitch contour, both Fónagy and Magdics and Freese and
Maynard observe a narrow pitch range with a slightly falling contour in reference to
softness/tenderness. Fónagy further reports rounded undulating fundamental frequency
curve in phrases expressing positive emotions, such as joy, tenderness and longing. In the
present study, we measured pitch level, as well as several pitch dynamics variables at the
syllable level. We hypothesize that softened phrases would exhibit gradual local changes
(moderate slope) within a bounded range. We had no specific prediction for the role of
pitch level in the perception of softness, as the literature is inconclusive in this respect.
We did not perform a quantitative analysis of the shape of intonation contours. We refer
the interested reader to Tsur ("Kubla Khan" Poetic Structure, Hypnotic Quality and
Cognitive Style: A Study in Mental, Vocal and Critical Performance) for a detailed
intonation analysis of the readings analyzed in the current study.

To summarize, ‘softened voice’ is an umbrella term describing the subjective
psychological perception associated with certain vocal manipulations. We hypothesize
that some subset of the following acoustic properties is sufficient (but not necessary) to
mediate the perception of softness: breathy voice, speech prolongation, lowered intensity
level, and gradual and bounded intensity and pitch contours.

**Research approach**

In studying the aesthetic qualities of sound, one may distinguish an exploratory-
qualitative, and a confirmatory-quantitative approach. The latter counts the number of
subjects who claim to have perceived a certain emotion in a text; the former offers
“plausible” hypotheses based on psychological and linguistic research to account for such
perceptions, even of a single participant. As Tsur (Playing by Ear and the Tip of the
Tongue: Precategorial Information in Poetry, 271-4) has demonstrated, rigorous
confirmatory-quantitative studies in reader response that are not preceded by proper

---
² We think that the two modes exploit different aspects of pitch. According to Ohala, aggressive animals
(mammals as well as birds) use low pitch, and appeasing animals use high pitch. Low pitch is produced
by large vocal tracts, high pitch by small vocal tracts. Animals use high pitch when they want to suggest
“I am small, I can do no harm, you may relax” — hence the tender effect. According to Ohala’s
experiments, the same intonation contour stripped of words is perceived as more dominant in a lower
register than in a higher register. On the other hand, higher pitch consists of faster vibrations; lowering
one’s pitch is in harmony with slower pace (cf. Tsur, Poetic Conventions as Cognitive Fossils: 201-206).
exploratory-qualitative analysis are prone to predetermine the results of the research. Gafni and Tsur present confirmatory-quantitative analysis of the emotional correlates of softened voice; in the present paper we present the underlying qualitative explorations of the physical properties of softened voice. Toward the end of this paper, we compare the results of this discussion to the statistical results in Gafni and Tsur.

Method

Material

As an attempt to find acoustic correlates that account for the perceived softened voice, we analyzed two readings of “Kubla Khan” (Coleridge, “Kubla Khan”) – one by Alex Jennings and one by Michael Sheen. The selection of these readings allowed us to study the realization of softened voice in different speakers. As to why we have chosen only two actors and why precisely these two actors, in his book on “Kubla Khan”, Tsur (“Kubla Khan” Poetic Structure, Hypnotic Quality and Cognitive Style: A Study in Mental, Vocal and Critical Performance) analyzes the readings of this poem by four actors, from the point of view of rhythmical performance. Since we are concerned here with microtextures, we were compelled to confine ourselves to the analysis of relatively few examples. We have chosen those two actors from our initial corpus, whose comparison yielded the most significant results. We didn’t intend to provide a definitive study covering all possibilities, but only a pilot study establishing the basic dynamics.

The recordings were taken from commercially-available, natural poetry recitals of leading British actors that were not conducted for the purpose of scientific investigation. Therefore, the studied material was not the result of a calculated process that aimed to produce vocal expressions of specified emotions in a controlled environment. However, our line of investigation focuses on studying the acoustic properties of speech samples given the generated subjective perceptions. By making the perceptual factor fixed (to some degree), the variable nature of the acoustic material becomes less of a concern. Moreover, it has been argued that natural emotive speech may possess subtle acoustic features that cannot be generated easily under controlled conditions (e.g., Scherer, “Vocal Communication of Emotion: A Review of Research Paradigms”). The following anecdote may illustrate this point. A few years ago, the second author of this paper sent to a leading British actor a paper analyzing a recording of his. The actor said he was overwhelmed by the thorough attention he had received, but added “Had I known that my reading would
be scrutinized in such a way, I would not have been able to utter a single word”. Thus, even if we could afford to engage professional actors from the London theatres and hire His Master’s Voice’s recording studios, we could not ensure that their performances would be optimal, only that the recordings satisfied certain technical prescriptions. In fact, we were not interested in the actors’ intended effect, from which the above considerations ought to be kept separate. For our results we needed only two sets of information: the recording’s perceived effect, and the sound processor’s output, without distinguishing between the actor’s intention and the recording conditions’ contribution to the effect.

Listen to four excerpts from “Kubla Khan”, read by Jennings and Sheen. These samples provide some examples of softened voice in both readings, as well as a contrast between a softened and a non-softened reading. The phrase Caves of ice appears twice in the poem, and we have detected the softening effect in both instances for both readers.

<table>
<thead>
<tr>
<th>Jennings (line 36)</th>
<th>Sheen (line 36)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jennings (line 47)</td>
<td>Sheen (line 47)</td>
</tr>
</tbody>
</table>

Similarly, both reciters seem to soften the phrase for he on honey-dew hath fed, though it appears to be much more pronounced in Jennings’s reading.

<table>
<thead>
<tr>
<th>Jennings</th>
<th>Sheen</th>
</tr>
</thead>
</table>

By contrast, Jennings reads the phrase I would build that dome in a most softened voice, while Sheen’s reading of that phrase is perceived as “hard”.

<table>
<thead>
<tr>
<th>Jennings</th>
<th>Sheen</th>
</tr>
</thead>
</table>

Procedure

We manually annotated each of the recordings, as exemplified in Figure 2 (The annotation and analysis were conducted using the speech analysis software PRAAT (Boersma and Weenink)). The recordings were annotated both at the segmental and syllable level. To be sure, we analyzed not only our samples of interest, but the complete recitals from which the softened samples were taken. Note that softened voice is not a strategy applied to the entire poem, but rather a special effect used by the readers to mark certain phrases in certain contexts. Thus, only a subset of the performance can by characterized as softened. Despite the relative scarcity of the effect, we managed to gain a reasonable
amount of data which contrasted “softened” and “non-softened” phrases, and was amenable for statistical analysis.

The second author of this paper, who has conceptualized the notion of softened voice, labelled phrases as either soft, or non-soft, according to his intuitions. Then, the first author measured several acoustic parameters for each of the annotated segments and syllables (see ‘Acoustic analysis’ section below). The acoustic analysis was performed independently of the labelling to reduce potential bias in the selection of the data. Softened and non-softened words were compared to determine whether they were different with respect to certain acoustic properties (see ‘Statistical analysis’ section below).

Acoustic analysis

We measured several acoustic parameters in each annotated segment and syllable. Parameters measured in individual segments were intended to detect breathy voice. According to the vast literature on breathy voice (e.g., Wayland and Jongman), this quality can be manifested in various ways by different speakers. Therefore, we measured several acoustic parameters known as possible cues for breathy voice.

It is often noticed that the first harmonic of a breathy vowel has a larger amplitude than the second harmonic (Figure 1). Thus, for every uttered vowel, we measured $H_1 - H_2$, the difference between these amplitudes.
The second category of acoustic correlates of breathy voice is *spectral slope*. The spectrum of breathy vowels often exhibits a steeper slope compared to the spectrum of non-breathy vowels. That is, the amplitude of higher overtones drops more steeply with the increase of frequency in breathy than in non-breathy vowels. Here we used three slope measures. *Spectral emphasis* is the amount of acoustic energy stored in high overtones (e.g., those that have a frequency of 1.5 times the fundamental frequency, or higher; Traunmüller and Eriksson). In an unaspirated, breathy voice, the higher overtones contain a smaller amount of energy compared to a non-breathy sound, thus yielding a lower spectral emphasis, an indication of a steep spectral slope. An aspirated breathy vowel contains relatively large amounts of energy in the high frequency band, and therefore its spectral emphasis would be closer to (and even higher than) that of a non-breathy vowel.

Another measure of spectral slope is the *Hammarberg Index* (Hammarberg et al.), the difference between the energy of the highest spectral peak in the 0-2000 Hz range and the energy of the highest spectral peak in the 2000-5000 Hz range. The last measure of spectral slope used in this study is the ratio of energies of high-frequency spectral band (4000 Hz and higher) to low-frequency spectral band (4000 Hz and lower) (H/L; Hillenbrand, Cleveland, and Erickson). An unaspirated breathy vowel is expected to have lower H/L and higher Hammarberg index than a non-breathy vowel. An aspirated breathy vowel may have a similar or even higher H/L and a lower Hammarberg index than a non-breathy vowel.

We also used two acoustic measures to estimate periodicity. Autocorrelation (Hillenbrand, Cleveland, and Erickson) can be taken as indication of the accuracy at which periodic patterns in the sound repeat themselves exactly. *Harmonicity*, or *Harmonics-to-Noise Ratio (HNR)* measures the relative dominance of periodic to aperiodic (“noise”) components of the voice (see Boersma; Wayland and Jongman).

A breathy vowel is expected to have lower values on periodicity measures compared to a non-breathy vowel (HNR: Wayland, Gargash, and Longman; Autocorrelation: Hillenbrand and Houde). However, note that periodicity is affected by various factors. Thus, measurements of periodicity have a potential difficulty in discriminating breathy voice from other types of aperiodic phonation, such as creaky voice (see Blankenship, “The Time Course of Breathiness and Laryngealization in Vowels”). Moreover, it has been claimed that HNR is affected mainly by low-frequency spectral components (Hillenbrand et al.). Thus, HNR can, in fact, be higher in breathy vowels compared to clear vowels (Wayland and Jongman).
In addition to the segmental analysis, we also examined possible prosodic contribution to voice softening. For each syllable in the recorded readings, we measured duration, as well as several pitch- and intensity-related parameters: including mean value, normalized excursion (the ratio between the maximum and minimum values of intensity/pitch divided by syllable duration), and the mean absolute slope (the magnitude of amplitude difference between two adjacent time points, divided by the time difference, and averaged over syllable duration). In general, we expect softened syllables to have higher duration, similar or lower intensity and similar or lower slope and range (excursion) of intensity and pitch compared to non-softened syllables.

Statistical analysis

Since softness judgments were acquired for words, the statistical analysis was also carried out on words. However, because acoustic measurements were taken from individual vowels and syllables, the raw data had to be manipulated to be able to represent whole words. This was done as follows: first, each parameter measured on the segmental level was averaged by vowel quality (softened and non-softened vowels together). Thus, we obtained separate grand mean values of $H_1 - H_2$ for /i/ and /u/. These calculations were performed separately for each reader. Second, the grand mean values were subtracted from measures obtained for individual uttered vowels (e.g., the mean value of $H_1 - H_2$ over all the /i/s in one recording was subtracted from the value measured for each uttered /i/ in that recording). Thus, the data was transformed from raw values to deviations from mean values. Finally, to obtain word-level scores, the modified values obtained for all vowels in a word were averaged by vowel duration. Thus, we expected that a word containing one long, highly-breathy vowel and one short non-breathy vowel would be perceived as relatively breathy (and softened) due to the dominant breathy vowel. Similarly, we used duration syllable weighted averages to obtained word-level scores for the prosodic parameters.

To assess whether certain acoustic parameters contributed to the perception of softened voice, analysis of variance (ANOVA) was done on scores for “softened” and “non-softened” words. Multiple comparisons were corrected for according to a False Discovery Rate (FDR; Benjamini and Hochberg) of 5%.

The fact that the analysis is based on the judgments of a single listener suggests that the results cannot be assumed to hold across listeners. However, since the studied
effect is subjective and not widely-recognized (unlike basic emotions such as sad, or happy), there is no guarantee that judgments from a larger pool of listeners would produce consistent results. Instead of aiming for generalizability, our immediate goal was to assess self-consistency. In other words, in this case study we investigated whether the subjective perceptions of our single subject, which seem to be rather confident, had consistent acoustic correlates.

To further establish the psychological reality of the softened voice, we tested whether the acoustic parameters that significantly contributed to the perceptual judgments of our informant also correlated with emotive judgments provided by naïve listeners in a previous study of the same two recordings (Gafni and Tsur).

Results

Perceptual judgments

Table 1 lists the softened phrases in Jennings’s reading (37 out of the 355 words of the poem). Each phrase is identified by its line number in the poem.

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Line</th>
<th>Phrase</th>
<th>Line</th>
<th>Phrase</th>
<th>Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>earth</td>
<td>18</td>
<td>of</td>
<td>47</td>
<td>floating</td>
<td>50</td>
</tr>
<tr>
<td>chaffy</td>
<td>22</td>
<td>ice</td>
<td>47</td>
<td>hair</td>
<td>50</td>
</tr>
<tr>
<td>sunny</td>
<td>36</td>
<td>heard</td>
<td>48</td>
<td>circle</td>
<td>51</td>
</tr>
<tr>
<td>caves</td>
<td>36</td>
<td>see</td>
<td>48</td>
<td>thrice</td>
<td>51</td>
</tr>
<tr>
<td>of</td>
<td>36</td>
<td>there</td>
<td>48</td>
<td>eyes</td>
<td>52</td>
</tr>
<tr>
<td>ice</td>
<td>36</td>
<td>should</td>
<td>49</td>
<td>holy</td>
<td>52</td>
</tr>
<tr>
<td>damsel</td>
<td>37</td>
<td>cry</td>
<td>49</td>
<td>dread</td>
<td>52</td>
</tr>
<tr>
<td>Abyssinian</td>
<td>39</td>
<td>beware (1)</td>
<td>49</td>
<td>he</td>
<td>53</td>
</tr>
<tr>
<td>music</td>
<td>45</td>
<td>beware (2)</td>
<td>49</td>
<td>honey</td>
<td>53</td>
</tr>
<tr>
<td>build</td>
<td>46</td>
<td>his (1)</td>
<td>50</td>
<td>hath</td>
<td>53</td>
</tr>
<tr>
<td>dome</td>
<td>46</td>
<td>flashing</td>
<td>50</td>
<td>paradise</td>
<td>54</td>
</tr>
<tr>
<td>air</td>
<td>46</td>
<td>eyes</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>caves</td>
<td>47</td>
<td>his (2)</td>
<td>50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In what follows, we will analyze the acoustic properties of the above listed phrases, compared to the remaining phrases in this reading.

---

3 Here we discuss our main findings. The full results of the analysis can be found in an online supplementary file.
First, “softened” words in Jennings’s performance have significantly different values of $H_1-H_2$ ($F(1,353) = 99.77$, $p<10^{-20}$), Spectral emphasis ($F(1,353) = 31.05$, $p<10^{-7}$), and $H/L$ ($F(1,353) = 48.51$, $p<10^{-10}$) from his “non-softened” words. The softened words are characterized by positive $H_1-H_2$ deviation (5.35 dB) and negative spectral emphasis deviation (-2.67 dB), while the non-softened words have a small negative $H_1-H_2$ deviation (-0.53 dB) and small positive spectral emphasis deviation (0.23 dB). This means that vowels in softened phrases have higher-than-average $H_1-H_2$ and lower-than-average spectral emphasis. These are the expected characteristics of breathy voice – the former finding suggests a vowel spectrum dominated by the fundamental frequency, and the latter indicates a steep spectral slope.

To be sure, we also calculated the mean raw values of $H_1-H_2$ to verify that Jennings’s softened vowels are indeed breathy. A grand mean of 5.68 dB for softened vowels (i.e., a positive value) versus a grand mean of -1.12 dB for non-softened vowels (i.e., a negative value) confirms that breathy voice is characteristic of his softened words, while his general voice has a slight tint of creakiness.

The second measure of spectral slope, the Hammarberg index, missed statistical significance after correction for multiple comparisons ($F(1,353) = 4.51$, $p=0.03$), but its positive value for softened words indicates steep spectral slope, similarly to the result for spectral emphasis. As for $H/L$, we note that softened phrases have a positive deviation (0.10%), on average, which suggests a relatively high amount of energy in high frequencies, compared to non-softened phrases (-0.01%). This seems to work against the interpretation of the low spectral emphasis and high Hammarberg index. However, recall that a high degree of aspiration adds energy to the high frequency part of the spectrum. In fact, as one can hear, Jennings’s softened phrases are often heavily aspirated. (Listen to an example: “paradise”).

Finally, the periodicity measurements of autocorrelation and HNR were not significantly different between softened and non-softened words in Jennings’s reading. To summarize, we established that on the segmental level, perception of softened voice in Jennings’s reading is significantly correlated with breathy (and probably heavily-aspirated) voice. We continue by reviewing Jennings’s reading at the syllable level.

Several syllable-level parameters seem to be correlated with voice softening in Jennings’s reading. On average, his softened syllables are longer in duration ($F(1,353) = 29.51$, $p<10^{-6}$) and lower in intensity ($F(1,353) = 6.79$, $p<0.01$) compared to non-softened syllables. In addition, the intensity contour of softened words exhibits more gradual
changes ($F(1,353) = 24.06, p<10^{-5}$) in a narrower range ($F(1,353) = 19.29, p<10^{-4}$) compared to non-softened words. Finally, among the pitch-related parameters, only the mean excursion was significantly different between softened and non-softened words ($F(1,353) = 17.3, p<10^{-4}$). The pitch contour of softened words exhibits changes in a narrower range compared to non-softened words. Together with the results of the segmental analysis, we conclude that perception of softened voice in Jennings’s reading is manifested by both segmental and prosodic properties.

Next, we turn to examine the acoustic analysis of Sheen’s reading. Table 2 lists the softened phrases in Sheen’s reading (41 words out of total 355).

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Line</th>
<th>Phrase</th>
<th>Line</th>
<th>Phrase</th>
<th>Line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xanadu</td>
<td>1</td>
<td>far</td>
<td>29</td>
<td>of</td>
<td>36</td>
</tr>
<tr>
<td>Kubla</td>
<td>1</td>
<td>ancestral</td>
<td>30</td>
<td>ice</td>
<td>36</td>
</tr>
<tr>
<td>Khan</td>
<td>1</td>
<td>prophesying</td>
<td>30</td>
<td>damsel</td>
<td>37</td>
</tr>
<tr>
<td>stately</td>
<td>2</td>
<td>war</td>
<td>30</td>
<td>dulcimer</td>
<td>37</td>
</tr>
<tr>
<td>ran</td>
<td>3</td>
<td>shadow</td>
<td>31</td>
<td>Abyssinian</td>
<td>39</td>
</tr>
<tr>
<td>man</td>
<td>4</td>
<td>of (1)</td>
<td>31</td>
<td>dulcimer</td>
<td>40</td>
</tr>
<tr>
<td>sunless</td>
<td>5</td>
<td>of (2)</td>
<td>31</td>
<td>Abora</td>
<td>41</td>
</tr>
<tr>
<td>sunny</td>
<td>11</td>
<td>pleasure</td>
<td>31</td>
<td>caves</td>
<td>47</td>
</tr>
<tr>
<td>spots</td>
<td>11</td>
<td>on</td>
<td>32</td>
<td>of</td>
<td>47</td>
</tr>
<tr>
<td>greenery</td>
<td>11</td>
<td>heard</td>
<td>33</td>
<td>ice</td>
<td>47</td>
</tr>
<tr>
<td>this</td>
<td>17</td>
<td>mingled</td>
<td>33</td>
<td>beware</td>
<td>49</td>
</tr>
<tr>
<td>chasm</td>
<td>17</td>
<td>measure</td>
<td>33</td>
<td>honey</td>
<td>53</td>
</tr>
<tr>
<td>huge</td>
<td>21</td>
<td>miracle</td>
<td>35</td>
<td>fed</td>
<td>53</td>
</tr>
<tr>
<td>from</td>
<td>29</td>
<td>caves</td>
<td>36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sheen’s vowel acoustics suggests that for him, too, $H_1 - H_2$ is a significant predictor for softened voice ($F(1,353) = 14.02, p<10^{-3}$), and in the same direction as in Jennings’s reading (softened words: 2.19 dB, non-softened words: -0.24 dB). Similarly, the analysis of spectral emphasis together ($F(1,353) = 39.85, p<10^{-9}$) with Hammarberg Index ($F(1,353) = 41.53, p<10^{-9}$) reveals that his softened words have steeper spectral slope than his non-softened words. On the other hand, there was no effect for the third measure of spectral slope, $H/L$. Moreover, examination of the grand means of raw $H_1 - H_2$ values shows that Sheen produces a relatively breathy voice throughout his performance. Although his softened vowels are generally breathy, and have a higher average value of $H_1 - H_2$ (8.5 dB), his non-softened vowels also have a positive value on average (5.2 dB).
Thus, it seems that breathy voice, per se, does not suffice to distinguish Sheen’s softened words from his non-softened ones. Instead, measurements of periodicity may provide the solution. The average deviation from the mean of HNR is positive (1.65 dB) and significantly higher \(F(1,353) = 14.08, p<10^{-3}\) in softened words than in non-softened words (-0.41 dB). A similar effect was found for autocorrelation \(F(1,353) = 6.48, p<0.01\;\text{mean deviation:} \quad \text{softened: 0.03, non-softened: -0.01}\). These results suggest that Sheen’s softened voice is more periodic than his non-softened voice. Recall that breathy voice typically has a low degree of periodicity. However, this generalization is relevant only in comparison to modal voice. Since Sheen’s voice is generally breathy, we do not conclude that lower degree of periodicity should be linked with softness. Instead, it seems that high degree of periodicity indicates a low degree of “roughness” (see Martin, Fitch, and Wolfe), which is compatible with the perception of softness. Since HNR depends mainly on low frequency components, the result is compatible with the effects of spectral emphasis and the Hammarberg index. Thus, despite the evidence for highly breathy voice in Sheen’s softened words, it seems that the softening effect on the segmental level is more complex.

To complement the acoustic analysis for Sheen’s reading, we discuss the results for the syllable level. Sheen’s softened syllables are significantly prolonged compared to his non-softened syllables \(F(1,353) = 9.56, p<0.01\). In addition, all intensity parameters are significant. Softened voice is achieved in Sheen’s reading by means of reduced intensity \(F(1,353) = 12.97, p<10^{-3}\), with a stable contour \(F(1,353) = 12.26, p<10^{-3}\) bounded in a narrow range \(F(1,353) = 8.39, p<0.01\). There seems to be no relation between pitch level and voice softening in Sheen’s reading. However, softened voice is correlated with a narrow pitch range \(F(1,353) = 7.03, p<0.01\). Thus, both segmental and prosodic factors contribute to the perception of softened voice in Sheen’s reading.

Emotional judgments

The analyses described above focused on the judgments of a single listener (the second author of this paper). Though these results are enough for the purpose of demonstrating the psychological reality of the softened voice, it would be even more convincing if we could demonstrate similar results for additional judges. In what follows, we attempt to strengthen our conclusions by referring to previously collected judgments for the two readings analyzed here.
In an earlier work, Gafni and Tsur asked 41 native speakers of English to evaluate the state of mind reflected in phrases taken from Jennings’s and Sheen’s readings of “Kubla Khan”. Among other things, participants were asked to evaluate (on a scale of 1-7) the degree to which each of the uttered phrases could be described as soft, as well as other intuitively related terms such as contemplative, affectionate, and yearning. The results of this judgment task were in accordance with the researchers’ intuitions. It would be interesting, therefore, to examine whether these judgments are correlated with the acoustic cues measured in the present study.

Participants evaluated twenty phrases from each recital with respect to eleven describing terms. Here we focus on four out of the eleven terms (see Gafni and Tsur and the online supplementary file for this paper). For each describing term (e.g., soft), we obtained a list of the average judgment values given by the participants for that term for all tested phrases. In addition, for each acoustic parameter (e.g., spectral emphasis), we obtained a list of acoustic word scores for all phrases (see “Statistical analysis” section). Then, we calculated the correlation between judgments and acoustic word scores, for each combination of describing term and acoustic parameter.

The results indicate that the participants in the study of Gafni and Tsur tended to give high ratings for the terms soft, affectionate, contemplative, and yearning when evaluating phrases which had high $H_1 - H_2$ score (i.e., positive correlations: soft: $R=0.70$, $p<10^{-3}$; affectionate: $R=0.78$, $p<10^{-4}$; contemplative: $R=0.76$, $p<10^{-4}$; and yearning: $R=0.78$, $p<10^{-4}$) and low spectral emphasis score (i.e., negative correlation: soft: $R=-0.86$, $p<10^{-5}$; affectionate: $R=-0.82$, $p<10^{-5}$; contemplative: $R=-0.84$, $p<10^{-5}$; and yearning: $R=-0.75$, $p<10^{-3}$). Thus, the correlation analysis supports our claim regarding the involvement of breathy voice in Jennings’s softened phrases. None of the other acoustic parameters was significantly correlated with participants’ judgments.

Next, we analyze the correlations between the judgments collected by Gafni and Tsur for Sheen’s reading and the acoustic parameters measured for the evaluated phrases. Again, twenty phrases from the complete read poem were analyzed (the same phrases analyzed in Jennings’s reading). There is one striking difference between the analyses of the two readings. There was no correlation between measures of $H_1 - H_2$ and ratings collected for Sheen’s reading ($p>0.1$ for all terms), as opposed to the strong correlations found in the analysis of Jennings’s reading. However, this is not very surprising given Sheen’s tendency to produce breathy voice throughout his reading. Yet, we did find significant (and some near significant) correlations between the ratings of describing...
terms and measures of spectral slope (spectral emphasis and the Hammarberg index), which indicate that the listeners did notice some spectral differences between Sheen’s softened and non-softened phrases (soft: Spectral-emphasis: $R=-0.65$, $p<0.01$; Hammarberg index: $R=0.75$, $p<10^{-3}$; affectionate: Spectral-emphasis: $R=-0.54$, $p=0.01$ (#); Hammarberg index: $R=0.73$, $p<10^{-3}$; contemplative: Spectral-emphasis: $R=-0.48$, $p=0.03$ (#); Hammarberg index: $R=0.69$, $p<10^{-3}$; yearning: Hammarberg index: $R=0.59$, $p<0.01$). Near-significant, positive correlations of the judgments with HNR are also in accordance with results for spectral slope (soft: $R=0.46$, $p=0.04$ (#); affectionate: $R=0.53$, $p=0.02$ (#)). In addition, we also found significant correlations between the ratings of three of the describing terms and mean intensities of the phrases (soft: $R=-0.68$, $p<10^{-3}$; affectionate: $R=-0.64$, $p<0.01$; contemplative: $R=-0.72$, $p<10^{-3}$). This suggests that, at least for the participants in Gafni and Tsur, reduced intensity (i.e., quiet voice) was an important cue for softened voice in Sheen’s reading. Finally, ratings of describing terms in Sheen’s reading were not significantly correlated with pitch-related measures.

**Discussion**

The present study was an attempt to establish a connection between acoustic features and perceptual qualities in listening to vocal performance of poetry. The results confirmed the psychological reality of the perception of ‘softness’, and revealed that perception of ‘softness’ can be mediated by several acoustic cues. Acoustic correlates of breathy voice were found to be fundamental to the perception of softness, in the two performances. This was found both in our own perceptually-based (softened/not softened) classification and in the quantitative judgments of naïve listeners.

Nevertheless, a comparison of the two examined readings suggests a speaker-dependent acoustic profile of softened voice. In Jennings’s reading, breathy voice is contrastive to the creaky voice characterizing the larger part of his reading. Thus, any signs of breathy voice are enough to render his voice soft. By contrast, detecting acoustic correlates for breathy voice is not enough for distinguishing Sheen’s softened and non-softened phrases. Rather, Sheen’s softened breathy voice seems to be more periodic than his non-softened breathy voice. In addition, low intensity was found to be a significant

---

4 (#) indicates that the result had uncorrected p-value < 0.05 but missed significance level after correction for multiple comparisons.
cue for softness in Sheen’s reading. Again, this finding is supported by our own intuitions and by the judgments of the participants in the study of Gafni and Tsur.

Despite some common findings in the correlation analysis for judgments in the current and previous study, there are some effects found only for the judgments of the second author of this paper. In particular, syllable duration, pitch excursion and the dynamics measures of intensity were not significantly correlated with the judgments collected by Gafni and Tsur. Such discrepancies could be attributed to sample size. On the one hand, participants in Gafni and Tsur judged only 20 phrases per reading which might not have had enough acoustic variability to produce the missing effects. On the other hand, in the current study we collected judgments from a single participant on the complete readings (with consistent findings). It is possible that this particular judge was sensitive to different aspects than the participants in Gafni and Tsur.

In addition, a methodological difference between the two studies can also account for the diverging findings. In the current study, judgments were binary (soft/non-soft), while Gafni and Tsur used a 1-7 ordinal scale to collect judgments. Also, the criterion for judgment was different between the two studies. In the current study, judgments were based on perceptual qualities (and, perhaps, to some degree on emotional qualities), while in Gafni and Tsur, judgments were about mood or state-of-mind. The latter problem might not be as severe as it seems if we consider the fact that the researcher who provided the judgments also designed the judgments experiment. Thus, even when providing the perceptual judgments, that judge was aware of the “higher” psychological effects of the softened voice. Nevertheless, more data is needed to support or refute the judgments collected in the present study.

Furthermore, the acoustic model proposed in the present paper specifies several acoustic parameters as potential cues for softness. However, the relations among these parameters remains underdetermined. For example, we found that both readers tended to produce softened phrases in lower-than-average intensity. According to our theory, speech can be perceived as softened as long as it is not uttered with great intensity. Nevertheless, more research is needed to determine whether low intensity alone can render a phrase softened (e.g., even when the voice is creaky). Similarly, we also found that softened syllables were, on average, longer in duration (especially for Jennings), compared to non-softened syllables. We hypothesize that prolonged articulation may enhance the perception of other acoustic cues for softness, but it is not clear whether
prolongation can render a phrase softened on its own. Notwithstanding the above, the question whether softened voice can depend on a single acoustic feature is not critical.

We now turn to discuss the psychological effect of softened and breathy voice. Gafni and Tsur found that judgments of phrase softness were highly correlated with judgments about *contemplative, affectionate* and *yearning* states of mind. In the present study, we found that acoustic cues that were correlated with the perception of softness, also tended to correlate with the perception of *contemplative, affectionate* and *yearning* states of mind. These results demonstrate that the softened voice serves to generate a general psychological atmosphere. The exact impressions experienced in that atmosphere are determined by a combination of the vocal features, the semantics of the text, and personal traits of the listener. Nevertheless, impressions associated with softened voice seem to share some common features; they tend to be rather passive mental states with low-to-medium degrees of activity (in the sense used in models of componential analysis of emotions and attitudes, such as the *semantic differential* (Osgood)).

The literature provides additional mental states associated with breathy voice and soft voice. Fónagy and Magdics report that speakers convey *longing* in a breathy voice. This can be paralleled with our finding regarding the association of *yearning* and breathy voice. Furthermore, Airas and Alku indicate the involvement of breathy voice in the expression of *tenderness*, and Laver claims that breathy voice is used by English speakers to communicate *intimacy*. Both qualities are closely related to *softness* and *affection*. This is supported, to some extent, by the perception experiment of Gobl and Ní Chasaide. In many contexts (e.g., in relation to beloved people), all of the above terms convey a similar idea of positive attitude towards someone or something. In fact, the phrase *sigh for* creates a concrete link between the breathy voice quality and feelings of *affection, yearning* and *longing*.

The literature also provides evidence associating breathy and soft voice with some “negative” emotions and mental states. For example, Freese and Maynard indicate that the delivery of bad news is often done with a breathy voice, which has a softening and soothing effect. Similarly, speakers in the experiment of Laukkananen et al. communicated *sorrow* (or compassion) with a breathy voice. Even more relevant to our subject matter, Fónagy reports that listeners tend to perceive sentences, which were intended to convey *tenderness*, as conveying *sorrow*, and vice versa. Additional studies suggest that breathy voice is linked with *sadness* (Gobl and Ní Chasaide; Airas and Alku; Scherer, “Vocal Affect Expression: A Review and a Model for Future Research”).


All of the above emotions and mental states can be characterized as states of reduced mental energy. Thus, emotions such as sorrow and sadness can be considered tender emotions (Gafni and Tsur; Tsur, *Playing by Ear and the Tip of the Tongue: Precategorial Information in Poetry*). Such emotions are typically communicated through breathy voice, which is characterized by muscular hypotension. In other words, mental states of low energy are vocalized with low muscular energy. The resulting “incomplete phonation” may cause breathy voice to be perceived as softer than “full” phonation, and suggests a general psychological atmosphere of reduced energy. This general atmosphere is individuated by the contents as sadness or tenderness.

We close the discussion by demonstrating how the acoustic features discussed in this paper may cue the perception of softness in the analyzed readings. We refer to Tsur's (“*Kubla Khan* Poetic Structure, Hypnotic Quality and Cognitive Style: A Study in Mental, Vocal and Critical Performance”) analysis of *damsel* in Jennings’s reading (Listen to example). The first vowel /æ/ is uttered in breathy voice, giving the word a softened quality. In the present context, the breathy, “incomplete” phonation may suggest unspecified emotional arousal, or perhaps the incomplete existence of the unattainable object of yearning. This perception is further enforced by the exceptional duration of the word; not in the sequence /dæ/, but in the continuant consonants following it: /m/ lasts 181 ms, /z/—106 ms (in recorded version of this word in the Merriam-Webster Collegiate Dictionary—Audio Edition, /m/ lasts only 64 ms, /z/ only 59 ms). This lengthening causes the speaker to linger on the word longer than expected without significantly affecting the stressed vowel. The emotional tint of the vowel /æ/ lends to this lingering a peculiar hue, as if the speaker were clinging to an object, or clinging onto a pleasant memory. The lingering on these continuants can become an iconic representation of the clinging to an object or a memory. One of the meanings of “nostalgia” can be suggestive here: “a wistful or excessively sentimental yearning for return to or of some past period or irrecoverable condition”.

**Conclusions**

The current study combines a qualitative analysis of listener response with a quantitative structural analysis of poetry reading. We have shown the potential of such a methodical analysis to account for the role of the reciter in mediating the experience of poetry. Further studies of similar fashion can help increasing our understanding of the psychological
dimensions of vocal performance. In particular, the current work is a “case study”, in which an acoustic model is constructed based on the perceptions of a single listener. The consistent results demonstrate the potential of studying intra-listener subjective perceptions. In addition, we have shown that this subjective perception might be shared by others. More research is needed to determine the intra- and inter-listener consistency of subjective perceptions in listening to poetry reading.

In addition, quantitative studies of intonation patterns, pitch contour and intensity contour dynamics will be valuable for complementing the findings of the present study. Finally, the present study focused on performance of English poetry. It would be interesting to conduct similar studies in additional languages, including such languages that use phonation (e.g., breathy voice vs. modal voice) as a phonological contrast.

References


