TEMPO 77 (305) 60–71 © The Author(s), 2023. Published by Cambridge University Press
doi:10.1017/S0040298223000098

VOICE WITHOUT SPEAKER: HUMAN SPEECH SYNTHESIS IN ACOUSTIC INSTRUMENTAL CONTEXTS

Andrew Chen

Abstract: This article provides an overview of compositional practices that invoke representations of the human voice by acoustic means, without the presence of actual human speakers, and, in identifying the limitations and deficiencies of extant practice, proposes a new method or methodological basis for achieving the same phenomenon, to different ends. Clarence Barlow's seminal *synthrumentation* method forms a central point of discussion; when considered alongside the work of, among others, Mayuzumi, Grisey and Ablinger, it becomes clear that, in existing synthrumental (or instrumental synthesis) practices, the innate inharmonic components of target material for reproduction have largely been discarded or overlooked. After a brief overview of the basic linguistic and perceptual features of speech, a new method designed to evoke whispered speech, with inharmonicity in mind, is demonstrated.

Human forms are, as a perceptual phenomenon, unique; people generally possess a subconscious yet immediate tendency to recognise discrete other things that bear resemblance to themselves, even when obscured, and to perform this more innately than learnt pattern recognition. This is true both in the act of recognising other people, as one might do when unexpectedly hearing voices from down the hallway in a residential building, when first spotting a figure walking ahead in the distance along the same path on a foggy day or when erroneously recognising human forms in objects that are not human at all. The latter phenomenon, known as pareidolia,[1] forms a significant part of shared human perceptive experience, whether viewed as a component of basic social function or more experientially complex 'animism… superstition, and magical thinking'.[2] This automatic perception of human forms has become a topic of wider interest, re-examined in the digital age in the concept of the 'uncanny valley': in essence, a hypothesis that proposes that increases in fidelity of virtual, human-like representation (originally as in robotics and CGI but

[1] Pareidolia is sometimes also used to refer to the broader misrecognition of *any* discrete object where it does not exist.

[2] Barbara Ellison and Thomas Bey William, *Sonic Phantoms*, ed. Francisco López (New York and London: Bloomsbury, 2022), pp. 19–22.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

since in other fields) may result in greater 'personal disquiet and a sense of strangeness' for an interacting viewer;[3] these being but two examples of a broader palette of unique, readily accessible affective phenomena derived from mechanistic origins.

Our tendency to recognise 'the human', then, comes bundled together with a number of illusory or subversive phenomena, and so it is reasonable to assume that human forms' deployment and manipulation in artistic contexts might be of a similarly visceral or immediate quality in their resultant affect. Artists have long recognised this, or at least followed similar strands of thinking: examples from visual media are abundant in the surrealist and Dada movements of the early twentieth century, as exemplified in Dalí's *Galatea of the Spheres*, Ernst's *Eye of Silence* and Magritte's silhouette works; but they also occur far earlier, around 1600, as in Arcimboldo's *Reversible Head with Basket of Fruit* and Wenceslaus Hollar's *Landscape*, among others. Each of these variously presents human forms such as faces and bodies in an illusory way, abstracted to such a degree that they are subsumed by the remaining subject(s) of each composition, but still recognisable by way of the pareidolia phenomenon or similar.

Sonic media need not be any different and, to this end, numerous musical works throughout history have explored the affective potential of dislocated human sounds, and of voices in particular. This category includes the use of offstage choirs in works such as Debussy's *Pelléas et Mélisande*, Puccini's *Madama Butterfly* and Holst's *The Planets*, and the attenuation of the sounds of brass or woodwind instruments by the voice of the player in respectively William Dougherty's *Three Formants* and Brian Ferneyhough's *Unity Capsule*. It could be argued, too, that the well-known ensemble pairing of trombones and voices in European sacred music, particularly in the sixteenth and seventeenth centuries, fits as well, as a kind of mimetic reference. However, if we are to extend the analogy of pareidolia and the uncanny, and the affective constructions of the visual works mentioned before, where literal humans are not present, it would be more apt here to focus on discrete vocal sounds created without the physical presence of the human voice itself.

Audio recording and production technologies seem to suit this category well: after all, they allow for vocal playback long after a speaker or singer was present. In the history of music technology, audio manipulation methods seem to emerge concurrently with broader creative interests in timbre as a formal or material basis for sonic works: I refer here both to the seminal avant-garde works of Schaeffer, Stockhausen et al., who, among other things, processed vocal recordings via physical duplication and alteration of tapes, and also to the methods of digital forms of vocal audio production, particularly in popular genre contexts.[4] In popular music vocoders and, more recently, formant filters and wavetables have been used to colour a signal with the timbral characteristics of the voice from the 1970s onwards; bitcrushing and pitch-shifting are, conversely, used to render a raw vocal recording 'synth-like' through distortion, as in the chorus hook of Skrillex and Diplo's 'Where Are Ü Now'. More generally, a number of methods are available to alter the recorded voice beyond

[3] Marcus Cheetham, 'Editorial: The Uncanny Valley Hypothesis and beyond', *Frontiers in Psychology*, 8, no. 1738 (2017), p. 4.

[4] Simon Frith, 'Afterword', in *The Relentless Pursuit of Tone: Timbre in Popular Music*, ed. Robert Fink, Melinda Latour and Zachary Wallmark (New York: Oxford University Press, 2018), pp. 373–75.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

its corporeal possibilities: these include simple EQ, resonance filters and pitch-correction (including Auto-Tune),[5] and the employment of proprietary software such as Melodyne to reconstitute formants, attack and decay characteristics altogether. The diffusion of 'obscured' human voices in these ways has become a mainstay of electronic music practice and production; indeed, it could be argued that *any* kind of audio playback of human vocality, even minimally processed, would fit this 'voices without speaker' category.

The emergence of what might be termed 'virtual vocality' and its current abundance in electronic practices has not, however, resulted in similarly widespread attempts to simulate human vocality *without* electronic diffusion or intervention; there is a relative dearth, in compositional contexts, of discrete vocal evocations achieved by purely acoustic means. Some significant examples do exist, particularly in the works of Clarence Barlow and Peter Ablinger, that I will discuss later, but their relative scarcity and apparent methodological limits suggest areas of acoustic vocal synthesis yet unexplored, which in turn could be useful in the future for new-music practitioners. This article grows out of my own prior research[6] and aims to explore this category of compositional practice – that is, simulations of discrete human vocality in acoustic (that is, non-electronic) contexts – and propose a novel methodological basis for achieving virtual vocality within it.

Conceptual origins; Barlow's *synthrumentation*; similar harmonically derived syntheses

In compositional practice, the evocation of sounds 'external' to a given instrumental resource is not particularly unusual, and has been evident throughout history, from utilisations of 'extra-musical' sounds in ancient and/or traditional medicinal or spiritual practices[7] to more modern examples, such as Haydn's 'clocks', the 'nightingales' of Respighi's *I pini del Gianicolo* and Stravinsky's *Jeu du Rossignol mécanique*, the 'decapitation' in Berlioz's *Symphonie fantastique* and the 'machines' of Mosolov's *Zavod*.

However, significant instances of music that non-corporeally invoke the voice are comparatively recent, the first such example probably being the laughter orchestrally depicted in Ravel's *Daphnis et Chloé* (1912). During the 'Danse grotesque de Dorcon', in the latter sections of Part One, the laughter of a crowd is depicted by punctuated, dissonant grace-note-adjacent interjections in the winds, juxtaposed against fractured tremolo in the upper strings.[8] It is notable that Ravel chooses to do this instrumentally, when he already has an on-stage choir at his disposal; it is thus an intentionally affective choice, heightened by conscious abstraction – what Fillerup describes as a parodic

[5] For a more in-depth contextualisation of Auto-Tune's usage in popular music production practice and its connections with other methods of vocal modification mentioned here, see Catherine Provenzano, 'Auto-Tune, Labor, and the Pop-Music Voice', in *The Relentless Pursuit of Tone: Timbre in Popular Music*, ed. Robert Fink, Melinda Latour and Zachary Wallmark (New York: Oxford University Press, 2018), pp. 159–76.

[6] Andrew Chen, 'Synthrumentation, Revisited: Towards a New Method of Additively Synthesising Speech in Acoustic Instrumental Contexts' (master's thesis, Royal College of Music, 2022), doi: 10.24379/RCM.00002280.

[7] Curt Sachs, *The Rise of Music in the Ancient World, East and West* (New York: Dover, 2008), pp. 19–23; Michael H. Thaut, 'Music as Therapy in Early History', in *Music, Neurology and Neuroscience: Evolution, the Musical Brain, Medical Conditions, and Therapies*, ed. Eckart Altenmüller, Stanley Finger and François Boller (Amsterdam: Elsevier, 2015), pp. 146–49.

[8] Maurice Ravel, *Daphnis et Chloé* (Paris: Durand, 1913), p. 52.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

kind of 'literalism'[9] – and echoes the practices of those visual artists intending the 'subversion' of human forms in the examples I mentioned earlier. Ravel's vocal evocations are not composed with any systematic rigour, however: it is unlikely that he used audio recordings of laughter or transcriptions thereof as a basis for his orchestration, and the result is thus a kind of mimesis rather than a direct simulation of voices. They do, however, provide an important precedent and establish a general principle: that one must be initially selective in the choice of target material in order to recreate it effectively.

It is not until Clarence Barlow's *Im Januar am Nil* (1981) that we find the next major example of speech evocation in acoustic contexts, and one that is far more procedurally complex. Barlow devised a specific orchestration technique that he called "synthrumentation", which, by the transcription and selective reorchestration of self-generated vocal spectrograms, allows melodic lines to 'sound like speech'.[10] Barlow says of the technique, used in *Im Januar am Nil,* that:

> It is the string section which is synthrumentally treated with an underlying bass clarinet explicitly but softly playing the melody. The analysed sound material is a set of sentences in the German language excluding all phonemes containing noise spectra such as plosives and fricatives. In all a total of two hundred words were found based on the remaining – lateral, nasal and vowel – phonemes, out of which a number of 'meaningful' sentences were formed, e.g. An Müllmänner in Armenien nun ein Jahr lang erinnern (Now commemorate garbage collectors in Armenia for one year.[11]

In liner notes for the work's definitive recording by Ensemble Köln, he further writes:

> The timbre derives from the synthrumentation of nine concocted German sentences (e.g. one which contains the title itself: *Im Januar am Nil, Mumien* anmalen = 'In January at the Nile, painting mummies'). All text syllables are spectrally harmonic, comprised of vowel, approximant, liquid and nasal phonemes. Ideally the bowed 'words' should be comprehensible, but an ensemble of seven string instruments can only be approximative.[12]

Barlow used this technique in two later pieces, *Orchideæ Ordinariæ* (1989) and *Felle Hymnus van Verre* (2001). His resynthesis procedure is made clear in Figure 1, which schematically shows the overtones within a vocal spectrum to be transcribed to the instrumental forces of his ensemble.

Barlow's synthrumental technique is striking and effective in its aural results, creating a distinct impression of vocality in the ensemble, even though the timbres of its instruments are largely unaltered (and thus aurally identifiable). It is, however, as a matter of course, particularly reliant on an initial reduction of its target material, in this case to spectral harmonicity (that is, harmonic overtones). Most sounds contain inherent, inharmonic components that serve a significant role in establishing their perceptive character,[13] but Barlow's method omits

[9] Jessie Fillerup, 'Purloined Poetics: The Grotesque in the Music of Maurice Ravel' (Ph.D. thesis, University of Kansas, 2009), pp. 263–64.

[10] Clarence Barlow quoted in Stephan Kaske, 'A Conversation with Clarence Barlow', *Computer Music Journal*, 9, no. 1 (1985), p. 27. For more in-depth technical explanations of Barlow's technique, see: Tom Rojo Poller, 'Clarence Barlow's Technique of "Synthrumentation" and Its Use in *Im Januar am Nil*', *TEMPO*, 69, no. 271 (2015), pp. 7–23; and Clarence Barlow, liner note for *Musica Algorithmica*. 2018, Ensemble Modelo62, pp. 5–6. as well as various other Barlow writings.

[11] Clarence Barlow, 'On the Spectral Analysis of Speech for Subsequent Resynthesis by Acoustic Instruments', *Forum Phoneticum*, 66 (1998), p. 184.

[12] Barlow, liner note for *Musica Algorithmica*, pp. 5–6.

[13] Chen, 'Synthrumentation Revisited', pp. 24–25. 'Inharmonic' is used here to refer generally to the complementary set of non-harmonic frequencies in a given spectrum. Some

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

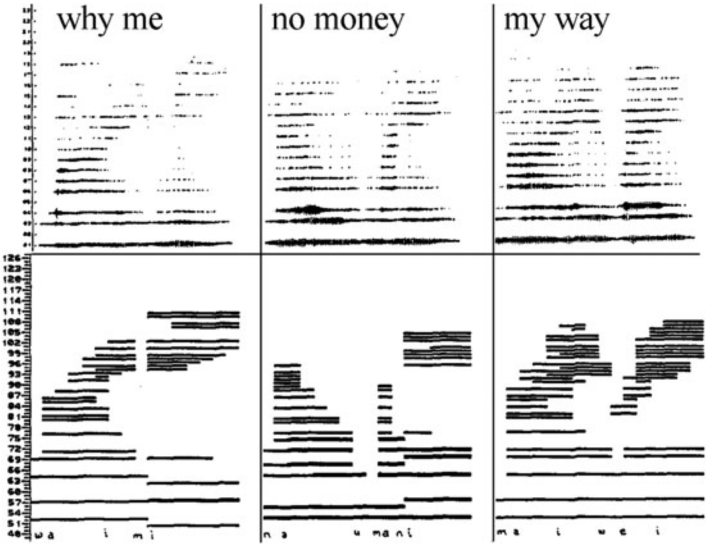


Figure 1:

Spectrograms of vocal recordings used in the synthrumental orchestration of Clarence Barlow's *Orchideæ Ordinariæ*. Reproduced by permission of Clarence Barlow.

from its target recordings any words that might contain significant noise components in the first instance; it also sidesteps the remaining presence of inharmonicity via the transcription process. This is both a methodological convenience as well as an intentional, artistic choice that allows for flexibility in the technique's application and variance in its resultant affect, Tom Rojo Poller identifying the synthrumental technique as 'one component among others' within a given work, effectively independent from melody, rhythm and form.[14] It could be argued, however, that it is a shortcoming of Barlow's method that it significantly underutilises the inharmonicity of vocal sounds.

There is a strong link between Barlow's synthrumentation and the material concepts of a number of earlier works. Of particular relevance here are Grisey's *Partiels* (1976) and Mayuzumi's earlier *Campanology I* (1957), both exemplars of a category sometimes referred to by scholars as works of instrumental re-synthesis or – the term I will continue to use here – synthrumentation.[15] Both are mixed acoustic ensemble works whose primary material derives from orchestrated transcriptions of the overtone spectra of target recordings (*Campanology I* being the first extant work to do so) and are distinct from Barlow's version of synthrumentation in that they treat such orchestrations as evolving sonic objects, rather than flexible modifiers of existing melodic components.

Much as in Barlow, however, these early synthrumental works involve a procedural reductivity that eliminates much of the inharmonic component of their target sounds. In the initial,

theorists, including Barlow himself, prefer to make a distinction between inharmonic overtones and noise spectra within this category.

[14] Poller, 'Clarence Barlow's Technique of "Synthrumentation"', pp. 9, 22.

[15] Nicholas Donin, 'Sonic Imprints: Instrumental Resynthesis in Contemporary Composition', in *Musical Listening in the Age of Technological Reproduction*, ed. Gianmario Borio (Burlington: Ashgate, 2015), p. 326. I am grateful to Julian Anderson for suggesting 'synthrumentation' to me as a useful, concise term to refer to any kind of sound resynthesis by acoustic means without the target source present, as it avoids any possible confusion between 'instrumental synthesis' or 're-synthesis' and other, more conventional, electronic-based methods; 'instrumental synthesis', for example, might also refer to an electronic synthesis *of* instrumental sound.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

'pre-compositional' working process of *Campanology*, which uses bonshō bells as a resynthesis target, Mayuzumi faithfully transcribed the microtonal pitch components of his spectral material (and was first attracted to bonshō in part because of their complex inharmonic character).[16] In later orchestral realisations he discarded them entirely, however, retrofitting the spectra into a 12EDO pitch space and nullifying much of their unique sonic character; he would later do the same in his *Mandala Symphony* (1960).[17] Grisey's *Partiels* is based at the outset on the spectrum of a trombone playing a loud, low E; the composer described the emergent processes of the piece as a kind of 'natural spectrum [drifting] with each repetition towards inharmonicity',[18] although such new frequencies are derived from a selective process of 'halation', as described by Krier,[19] rather than having anything to do with the fundamental inharmonic components of the trombone's sound (such as the sound of the breath, instrumental character, etc.). Initial pitch material in *Partiels* is harmonic, and the later inharmonicity of the work is a generative property rather than innate to its basic synthrumental material.

Common to all these works and other synthrumental attempts, then, is a reliance on the overtones of the harmonic series in the transcription and reorchestration process, and the intentional disregard of the inharmonic character of their target materials. A possible implication is that similar spectral transcription and reorchestration techniques could be devised and applied specifically with inharmonicity in mind; these might not only be of similar affect but also present different possibilities for orchestration and expressivity.

Adjacencies: Ablinger's talking piano pieces and Harvey's *Speakings*

Peter Ablinger has also explored the 'border area between abstract musical structure and [linguistic cognition]' via instrumental resynthesis,[20] particularly in his *Phonorealist* pieces, written since 1996. In these works there is an explicit link to the affective aims of the visual artists mentioned earlier; it is significant in this context that Ablinger has described himself as a 'painter [working] with compositional means'.[21] Like the synthrumental works I have discussed, Ablinger's phonorealist pieces were predicated upon acoustic resynthesis of target material and were performed live through acoustic instruments, but they notably excluded live human players entirely, opting instead for digitally assisted playback (through player-piano or similar instruments). Although the phonorealist pieces are thus not entirely acoustic, they still serve as an illustrative example of concurrent attempts to achieve vocal resynthesis without voices present.

As an example, Ablinger's *Deus Cantando* (2009), for player-piano, consists in its entirety of a literal resynthesis of a young person's

[16] Toshiro Mayuzumi quoted in Judith Ann Herd, 'The Neonationalist Movement: Origins of Japanese Contemporary Music', *Perspectives of New Music*, 27, no. 2 (1989), p. 137.

[17] Yuriko Takakura, 'A Visual Analysis Methodology for Music Compositional Process with Sound Resynthesis' (Ph.D. dissertation, Keio University, 2019), p. 51.

[18] François-Xavier Féron, 'Gérard Grisey: première section de *Partiels* (1975)', *Genesis*, 31 (2010), p. 93.

[19] Yves Krier, 'Partiels, de Gérard Grisey, manifestation d'une nouvelle esthétique', *Musurgia*, 7, nos 3–4 (2000), pp. 162–63.

[20] Peter Ablinger, 'Phonorealism', https://ablinger.mur.at/phonorealism.html (accessed 1 March 2023).

[21] Peter Ablinger quoted in Donin, 'Sonic Imprints', p. 326.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

German-language recitation of an International Environmental Criminal Court declaration.[22] Its compositional process is like synthrumentation in that it starts with target audio material intended for spectral rendering and reorchestration, but after that it diverges:[23] instead of relying on fast Fourier transform (FFT) analysis alone, Ablinger additionally incorporates constant-Q transform (CQT) and wavelet analysis, which allows for a broader, almost stochastic capturing of the target into precise individual pitches, durations and velocities; these are then approximated to 12EDO and transmitted via MIDI to the piano.[24] Other similar works include the *Quadraturen* series (since 1997) and *A Letter from Schoenberg* (2008); although these all differ in presentational contexts, and are often intended to be presented alongside various multimedia stimuli (such as the transcribed text), they utilise variations of the same basic technique.[25]

The sonic results of Ablinger's technique are notable for their immediacy and intelligibility as speech. Popular interest in 'speaking piano' illusions and their attendant effectiveness has been evidenced in the device's assimilation into internet meme culture, where it has become a focal point of a number of viral YouTube videos.[26] As an avenue for musical expression, however, phonorealism is somewhat inflexible: because of the extreme precision with which pitches are realised through the analytical process, it is not practicable for a live ensemble to perform the music, especially if they are to preserve the effectiveness of the resynthesis. This in turn presents barriers to the integration of the technique into broader musical contexts. Nevertheless, through its success in capturing speech's inharmonicity and thus realising its affectual aims, Ablinger's work suggests ways in which new methods might broaden their analytic scope beyond the harmonic limitations of synthrumentation and other earlier examples.

Another work that, although not strictly synthrumental, presents useful data for further methodological development is Jonathan Harvey's *Speakings* (2008), a mixed orchestral and live electronic work composed in association with Gilbert Nouno and IRCAM based on the concept of phonemic 'shape vocoding', the colouring of live orchestral instrumental sounds with identifiable vocal timbres.[27] The speech inflection of the work is effected both by acoustic means and live electronic intervention; the former was mostly produced using IRCAM's Orchidée orchestration software,[28] which is able to render the spectra of a user-selected 'SoundTarget' into a 24EDO pitch-space and generate orchestrations intended to acoustically resynthesise it.

Orchidée is highly sophisticated and does not significantly limit a practitioner's initial choice of target material, but it does present a number of procedural barriers for users: in particular, there is a finite set of possible orchestral sounds and techniques available in its inventory, and there is an inherent inflexibility in its use because of its functions' exclusive availability through a set user interface (UI). This

[22] G. Douglas Barrett, 'Window Piece: Seeing and Hearing the Music of Peter Ablinger', https://ablinger.mur.at/docs/barrett_window_piece.pdf (accessed 1 March 2023), p. 5.

[23] Ibid.

[24] Winfried Ritsch, 'Robotic Piano Player Making Pianos Talk', *Proceedings of the 8th Sound and Music Computing Conference, Padova* (2011), pp. 395–396.

[25] Robert Hasegawa, 'Timbre as Harmony – Harmony as Timbre', in *Oxford Handbook of Timbre*, ed. Emily I. Dolan and Alexander Rehding (Oxford: OUP, 2021), p. 537.

[26] 'Auditory Illusions: Hearing Lyrics Where There Are None', www.youtube.com/watch?v=ZY6h3pKqYI0 (accessed 28 February 2023); and 'Shrek but the ENTIRE MOVIE is converted to MIDI', www.youtube.com/watch?v=wcehaxidJZk (accessed 28 February 2023).

[27] Jonathan Harvey, *Speakings* (London: Faber, 2008), p. iv.

[28] The Orchidée software was named by IRCAM after Clarence Barlow's *Orchideæ Ordinariæ*.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

could present an opaque inflexibility for composers wishing to refine their orchestrations and the efficacy of their resyntheses. In *Speakings*, whose acoustic resyntheses were primarily derived from Orchidée, Nouno noted that much 'interesting structural information [was] seemingly lost' through the software's automated process; in fact, the live electronic component of the piece, and its highly specific physical and digital set-up components, explicitly aims to compensate for this, and to heighten the virtual speech's illusory effect.[29] Nevertheless, *Speakings* is a highly expressive and convincing musical work that successful achieves its intended illusions; importantly, too, it reinforces my proposal that broadening the analytical scope of working methods can increase the efficacy of a vocal evocation.

The inharmonicity of speech and its perception

In my view, it is clear that any development of synthrumental practices should maintain some degree of flexibility in working procedures and/or sounding results in order to both achieve its desired affect and offer creative utility for a practitioner. At the same time, it should also interrogate inharmonicity, a component of target sounds that has so far largely been disregarded.

Speech essentially functions as the filtering of frequency excitations through the acoustic cavity of the supralaryngeal tract – primarily, the mouth – itself an 'encoder of speech sounds'.[30] In creating excitations, usually through the larynx or through aspiration, humans configure their mouths into diverse but precise shapes, which allow for the production of phonemes (individual units of speech) and in turn the components of words and intelligible communication. From a spectral perspective, this shaping of the mouth manipulates the width and position of a number of frequency regions where filtering is less prominent and through which excited frequencies more readily pass; these are known as formants and are crucial in the differentiation of different phonemes, particularly vowel sounds.[31] The number of formant regions in vocal production is theoretically infinite, although in practice usually only the lowest frequency regions – labelled F1, F2, F3, and F4, in ascending order – are of analytical use.

Importantly, the frequency ranges of formants are not linked in any way to the harmonic series or overtones in and of themselves. Rather, it is in phonated speech – where the larynx is activated and a fundamental tone is produced – that overtones coincide with these formant regions and, as a result, these are clearly visible on spectrograms, hence the emphasis on harmonics in Barlow's method. Conversely, frequencies filtered through formants can be entirely inharmonic and yet intelligible as speech, as is the case in whispering, where broad spectrum excitation is caused by exhalation through the lungs (see Figure 2).[32]

Formants are not the only inharmonic component of speech, nor the only such component crucial to speech's intelligibility; other acoustic factors include transients (as seen in plosives or stops),

[29] Gilbert Nouno, Arshia Cont, Grégoire Carpentier and Jonathan Harvey, 'Making an Orchestra Speak', *Proceedings of the 6th Sound and Music Computing Conference, Porto* (2009), p. 280.

[30] Ville Pulkki and Matti Karjalainen, *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics* (Chichester: Wiley, 2015), pp. 90–91.

[31] Chen, 'Synthrumentation Revisited', pp. 37–43.

[32] Ibid., pp. 44–47.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

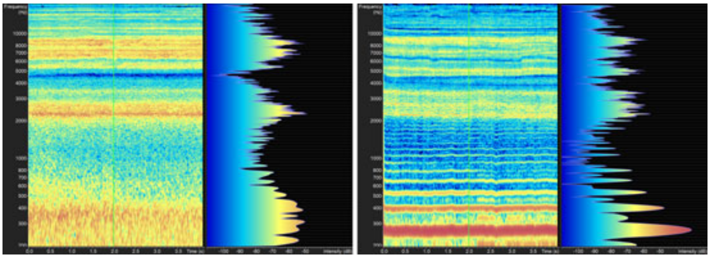


Figure 2:
On the left is a phonated [i] phoneme, on the right a whispered [i] phoneme, both normalised to peak amplitude of –24 dB; in both the regions of heightened volume are similar but include different discrete frequencies within.

other sonic residuals (including the breath) and broader linguistic structural features such as prosody and coarticulation.[33] In the context of devising a new synthrumental method, however, the inharmonic filtering component of formant regions presents itself as an attractive area of interest and one largely untapped so far.

Additive resynthesising of whispered speech in acoustic contexts

I want now to propose a new method, or methodological basis, for additively synthesising speech sounds in acoustic instrumental contexts, with whispered speech as a discrete target sound. What follows is both an abridgement and refinement of my own previous research.[34]

The excited frequencies of whispered speech occupy entire formant bandwidths, rather than just harmonic peaks of a given fundamental that coincide with them; thus some kind of acoustic instrumental realisation of a similarly inharmonic and wide-bandwidth spectral character could recreate them with a reasonable degree of fidelity. Damping (sometimes referred to as 'rauschen'), a technique available on all bowed string instruments, is highly effective for this purpose: by lightly muting a string with multiple fingers of the left hand and bowing flautando and molto sul tasto, a player can create a wispy, white-noise-like sound, with a faint residue of pitch (see Figure 3). The technique is most readily accessible on the outer strings of a given instrument, especially when playing high on the fingerboard. Much as in whispered speech, spectrographic representations of this damped sound show a wide band of excited frequencies around a central peak, and at relatively low dynamics (see Figure 4).[35]

This bandwidth effect is heightened when the playing technique is massed and clustered by a number of string instruments, as one might find in an orchestral section; unlike in conventional tone clusters, the net effect is of a single sonic mass rather than an ebbing stream of individually identifiable pitches. My prior live experiments suggest that players will readily, intuitively divide pitch choices within a given specific intervallic range to create perceptually even clusters, without the need for individual microtonal note assignment.[36] In order to achieve the effect, recordings of whispered speech can be transcribed spectrographically and redeployed, much the same as in Barlow's synthrumentation, with the important distinction that banded formant regions are reproduced, rather than singular, discrete pitches as in

[33] Jonathan Harrington and Steve Cassidy, *Techniques in Speech Acoustics. Vol. 8. Text, Speech and Language Technology* (Dordrecht: Springer Netherlands, 1999), pp. 1–4.
[34] Chen, 'Synthrumentation Revisited', pp. 51–62.
[35] Ibid., pp. 55–59.
[36] Ibid.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

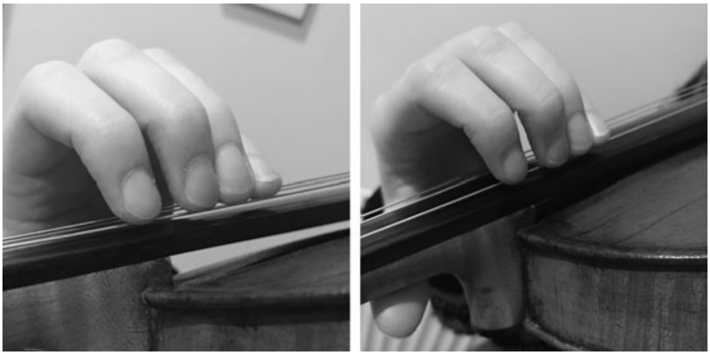


Figure 3:
Photograph of a violinist's left hand performing the damping technique on the G string. Multiple fingers damp the string without stopping it completely so that the resonance is muted and no harmonics are sounded. In this photo it is the position of the fourth finger that determines the sounding pitch.

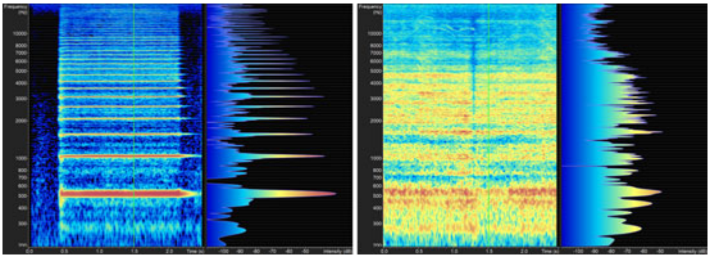


Figure 4:
Spectrograms: on the left is a violin playing a C ordinario, on the right the same violin playing the same C with the damping technique, both normalised to a peak amplitude of −24 dB. Note the relative lack of distinct peaks in the damped example.

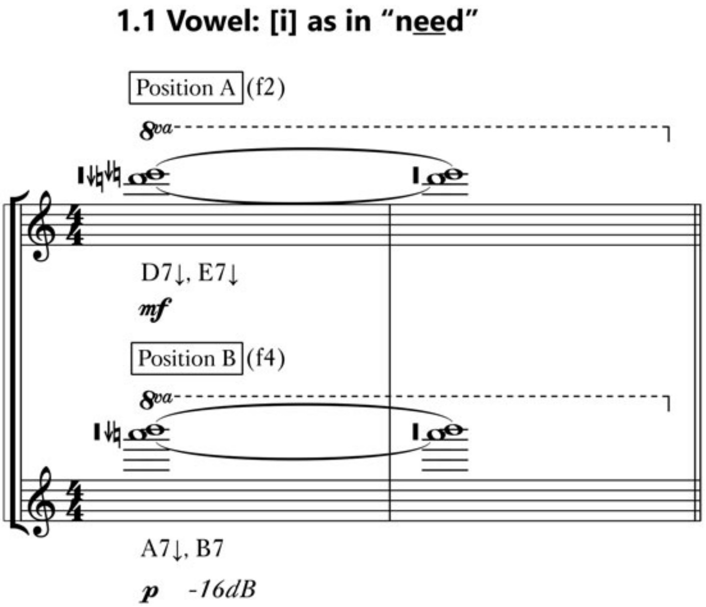
Im Januar am Nil. Clustered damped notes can be readily assigned to formants in much the same way that they are grouped by linguists (F1, F2, F3, etc.). The minimum number of string instruments required is determined by the largest interval within which a cluster is to be created (usually corresponding to the lowest formant bandwidth of the target synthesis material); in my own experiments the total number of instruments used has never exceeded 12, although this is by no means a prohibitive restriction. Example 1 shows a possible notation.

For this article, I have created three new pairs of whispered phoneme targets and damped cluster synthrumental realisations on string instruments; their orchestration follows a similar procedure to the one described above, and I have provided spectrograms of each for visual comparison (see Figure 5a–c). In this case, the examples were created by multi-tracking several recordings of a violin; a live realisation of this technique can be heard at https://researchonline.rcm.ac.uk/id/eprint/2280.[37]

The spectrograms of each 'target-synthrumentation' pair are evidently similar, particularly in their discrete formant regions and distribution of noise-like frequency content. Any more conclusive judgement on their perceptibility as 'speech-like' is beyond the scope of this article, but they are from my own standpoint aurally convincing.

The method is not flawless, however, as the spectrograms show a number of complicating factors, in this case mostly to do with the innate timbral characteristics of the violin; similar to Barlow's *Im Januar am Nil*, it is impossible to entirely escape the characteristics

[37] These prior examples are different from those shown in this article.



Example 1:

A notational realisation of the indeterminate cluster technique; the vertical bar to the left of the noteheads and accidentals indicates that a playing section should distribute notes within the written interval, as previously described. Additional annotation is for mnemonic purposes only.

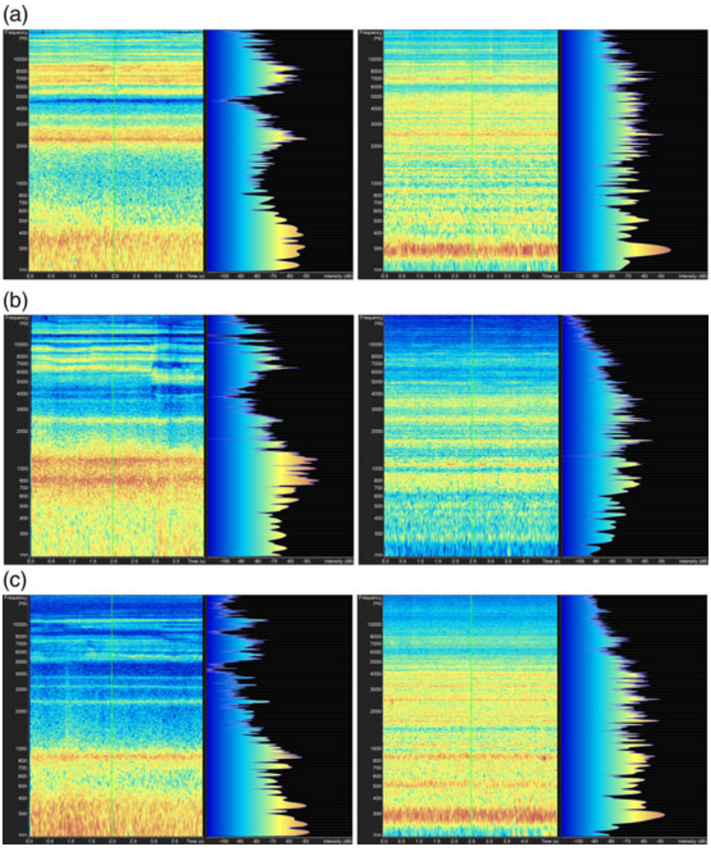


Figure 5

(a–c): Three pairs of whispered phonemes and their synthrumentally realised counterparts (using the method described). All recordings are normalised to −24 dB. Whispered [i] (left) versus newly synthrumented [i] (right). Whispered [a] versus newly synthrumented [a] (right). Whispered [u] versus newly synthrumented [u] (right).

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press

of the instruments for which a synthrumentation is written. Furthermore, there are still some instances where thin, clear tones are spectrogrammically visible, such as in the orchestrated [u] phoneme, where there is a lack of blending between parts; this can be tempered quite readily, however, as the smoother bandwidths of the [i] and [a] synthrumental phonemes demonstrate. More pertinently, there is significant extraneous frequency content (for example, within the 3000–4000 Hz range in the [i] synthrumentation) that was not present either in the original target recordings or in the performance instructions. These additional frequencies are a by-product of the instruments themselves, a result both of their own characteristic formant regions and residual noises from contact between bow and string. Perhaps these additional frequencies could be attenuated through variations of technique, such as changes of bow contact point and speed, or by using different instruments entirely: indeterminate damped cluster orchestrations are possible with any bowed strings, after all, although these would possess different acoustic characteristics.

In any case, as a compositional device, rather than purely as a theoretical construct, this novel method of synthrumentation not only enables a procedure for vocal resynthesis in a purely acoustic instrumental context, but can also be flexibly integrated into different musical contexts. For example, it need not remain a static, sonic object or effect: it is easy to make transitions in and out of the damped technique by gradually changing left-hand pressure, with the resultant effect of discrete ordinario tones emerging from a synthrumentation like distinct shapes in mist. Nor does this method restrict other aspects of musical form and expression, as is perhaps the case in Ablinger's work, suggesting that there is much potential for creative interplay and experimentation when applied within different musical contexts.

My intention has been to explore the ways in which acoustic additive synthesis can realise the inharmonicity of speech, largely overlooked in earlier attempts. My proposed method follows Barlow's in its reliance on FFT-derived spectrograms rather than other analytical methods, such as those used by Ablinger or Harvey, which may themselves also be useful for novel synthrumental generation. However, to make such a judgement, or any broader judgement on either the efficacy of vocal resynthesis in acoustic contexts or the philosophical implications of its historical and present usage, is not my aim here. These are compelling areas of inquiry, but they require much wider, subsequent research; I hope instead that this article may be useful for those who share an interest in this topic and wish to utilise or develop aspects of these findings in their own work.

https://doi.org/10.1017/S0040298223000098 Published online by Cambridge University Press