Theo van Leeuwen*

Three sound bites: avenues for research in the study of speech, music, and other sounds

https://doi.org/10.1515/jwl-2025-0037
Received August 20, 2024; accepted September 15, 2025; published online October 20, 2025

Abstract: This paper argues for a multimodal 'phonology', a 'phonology' which returns to the original meaning of the word 'phone' as referring to 'voice' as well as 'sound' in general, and which can be applied to speech, music, as well as other sounds. Three areas are discussed. The first is rhythm. Systemic functional linguistics has reduced the role of rhythm to the 'foot', but others see rhythm as also playing a significant role at higher ranks and as the fundamental organising principle of all time-based multimodal texts. The second is voice quality or, more generally, sound quality, as a key resource for expressing identity as well as fleeting states and emotions. Based as it is on qualities common to all sounds, this too applies to all sound-based semiotic modes. The third is the meaning-making potential of pitch contours, which, the paper argues, primarily contribute to the communication of emotion – again in speech, music, as well as other sounds.

Keywords: multimodality; phonology; pitch contours; rhythm; sound quality

1 Introduction

In this paper I bring together aspects from earlier work on speech, sound, and music (e.g. van Leeuwen 1992, 1999, 2009) to argue for a *multimodal* 'phonology' – a 'phonology' which returns to the original meaning of the word 'phone' as referring to 'voice' as well as 'sound' in general, and which can be applied to speech, music as well as other sounds.

Three areas are particularly relevant for such a multimodal phonology. The first is rhythm. Systemic functional linguistics has reduced the role of rhythm to the 'foot', but others see rhythm as also playing a significant meaning-making role at higher ranks, and as the fundamental organising principle of all time-based multimodal texts (in this paper I will focus on sound, rather than also on body movement, but see e.g. Bolinger [1986], on the relation between intonation and gesture). The second is voice quality, or, more generally, sound quality, as a key resource for expressing

***Corresponding author: Theo van Leeuwen**, Department of Language and Culture, University of Southern Denmark, Odense, Denmark, E-mail: leeuwen@sdu.dk

identity as well as fleeting states and emotions. Based as it is on qualities common to all sounds, this too applies to all sound-based semiotic modes. The third is the meaning-making potential of pitch contours, which, I argue, primarily contribute to the communication of emotion – again, in speech, music, as well as in other sounds.

2 Rhythm and the structure of time-based multimodal texts

Systemic functional intonation theory focuses on the structuring of information. 'Tonality' demarcates boundaries between information units, 'tonicity' marks the focal point in each information unit, and 'tone' marks whether an information unit is (some kind of) assertion or query. Rhythm plays a minor role. Halliday (1985: 48), while recognizing that rhythm "imposes organisation on the sounds of language, particularly the patterning of syllables", argues that rhythm "does not by itself express contrasts in meaning".

However, information structure can also be realized in monotone speech, or in whispered speech, where tone is absent and pitch substituted by intensity, duration and vowel quality (cf. e.g. Meyer-Eppler 1957). In music, too, we can "eliminate variations in pitch, loudness, or timbre and still recognize the melody. But if we destroy the internal temporal relationships without disturbing the other variables, the melody becomes at once unrecognizable" (Lenneberg 1967: 108). And while pitch rises and falls are absent in monotone and whispered speech, rhythm is not. It can, on its own, realize phrasing and mark focal points. Only the finality or continuity of the phrases in monotone speech is realized by pitch ('low' and 'not low', Brown 1977). In monotone music, too, only the last notes of phrases move down (or sometimes up), from early Gregorian psalters to specific uses of monotony, as in the character of Death in Schubert's 'Death and the Maiden', or in Jobim's famous bossa nova tune 'One Note Samba'.

This insight is not new. It can for instance be found in Pike (1945, 1962) and in Daneš (1960), who stated that "rhythm in English is not just something extra, it is the guide to the structure of information in the spoken message" (Brown 1977: 43). In my own work (van Leeuwen 1999, 2005, 2025), I have extended it to other time-based modes. Rhythm is multimodal, structured the same way in every time-based semiotic mode. As Couper-Kühlen (1993: 11) has said, in a linguistic account of rhythm: "The principles of organization of speech, verse and music are surprisingly similar".

This also means that rhythm demarcates information units at different ranks, not just that of the foot and the phrase, the rhythmic equivalent of the 'tone group'. Pike (1962) recognized the 'simple rhythmic unit', the 'complex rhythmic unit', the

'phonological paragraph', and the 'rhetorical period'. Daneš (1960) distinguished the 'foot', the 'utterance section', and the 'utterance'. Martinec (2000, 2002) recognized seven levels, each consisting of a wave and a transition, with the end of each wave marked by a boundary tone (high, mid, or low, as measured relative to the baseline of unaccented syllables). In my own work (van Leeuwen 1982, 1992), I recognize the 'measure' (as another term for the foot), the 'phrase', the move, and the 'phase'.

Here is an example from my work on radio announcing speech. I should add that, in my view, rhythm should be analysed in an *embodied* way, through 'finding' the rhythm by bodily responding to it, for instance by tapping on the accented syllables or notes – and only then notating the rhythmic accents and boundaries found in this way in some kind of transcript, as I have done in the examples below. This is because the isochrony of rhythmic feet is perceptual – our subjective perception of isochrony compensates for minor differences in duration, and therefore does not necessarily correspond to 'objective' isochrony, as measured by a stopwatch or other similar instrument (Lehiste 1977). In the example below, the rhythmic accents are italicized and bolded, with the nuclear rhythmic accents (realized by an increase in intensity and/or duration and/or pitch) also capitalized. The boundaries between measures are indicated by single slashes, while phrases are placed between brackets and separated by double slashes, and 'moves' by double brackets and triple slashes.

> [[***I've*** read**/ *news*** at all/ ***SORTS*** of**/ *pla***ces and the the//] [***blokes*** that/ ***write*** the news/ ***HERE***//] [***EM***//] [***CER***tainly//] [***far/ EA***sier to/ ***read***//] [and that's/ ***not*** just because I'm/ ***wor***king here/ ***NOW***//] [than/ ***A***ny news I've/ ***read*** anywhere/ ***else***///]]
> [[the con/ ***tai***nership/ ***A***sian Re/ ***NOWN***//] [is/ ***due*** to leave/ ***BRIS***bane to**/ *day***//] [with a con/ ***sign***ment of u/ ***ra***nium/ ***YE***llowcake///]]

The first transcript comes from a research interview with a newsreader from a Sydney commercial radio station. I had asked him whether the news in commercial radio stations is more informal than that of the ABC, Australia's national public broadcaster, where he had also worked. His answer, a self-contained 'move' in the overall question-and-answer interview genre, consisted of seven phrases. The end of the move was indicated by a pause as he waited for my reaction or next question. The nuclear accents conveyed crucial information in each of the phrases. In the first phrase, the newsreader established his credentials for answering the question by emphasizing his extensive experience (he has worked at all **SORTS** of places). He then narrowed the focus to his current place of work (**HERE**) and took time to think (**EM**) before emphasizing that he would give a definite and deeply felt answer (**CER**tainly), namely that his station's news is **EA**sier to tread than **A**ny other news.

Thus, each rhythmic phrase was a meaningful step in formulating, on the spot, and step by step, the thinking process that built up his answer.

The second example transcribes how (part of) a news item was read by the same newsreader. The item as a whole contained two moves. The example shows the 'lead' move, which encapsulates the core information of the item, the departure of a ship transporting uranium. The second move described a previous event that made the ship's departure newsworthy, a protest demonstration that tried to stop it from loading the uranium. Here the phrasing did not result from formulating an idea on the spot but from the well-established style of professional news readers (van Leeuwen 1992). The tempo was even, the rhythm regular with all rhythmic feet except the phrase-final feet containing three syllables and the phrase boundaries neatly coincided with grammatical boundaries. The pauses between the phrases were quite long, as if to emphasize that each phrase contained a distinct item of information.

Experimental research (Abrams and Bever 1969) has revealed the role of rhythmic segmentation in 'processing' information. At phrase boundaries, we somehow collate the information of the preceding phrase and store that 'collation' in memory. At move boundaries, we do the same at a higher level. At the end of a news item, or a move in conversation, most of us will not be able to remember what was said word for word, unless a phrase is repeated several times. But we nevertheless end up with enough of a holistic, integrated understating of what was said to allow us to respond appropriately. Processing information over time is a dynamic embedding of the past into the unfolding present – and into the immediate future, as rhythm also allows us to anticipate the accented moments that will carry further important information.

Rhythm is also the key to understanding how different modes interact in time-based multimodal texts such as films. In a handbook for film and video editors, Karen Pearlman (2009: 17) describes how film editors respond in an embodied way to the rhythms of the filmed dialogue and actions.

> What editors do is to tune themselves to the rhythm of the material, drawing on their own experience of the rhythms of, for example, blinking. This knowledge of blinking rhythms they have perceived is implicitly compared to the rhythms they see in the rushes and cuts they are working on. As they continue to refine the cuts, they use their kinaesthetic empathy to relay the external rhythms which they perceive in the developing edits, *through* their internal rhythms, to create the rhythm of the film. (Pearlman 2009: 17)

One of the different modes in time-based multimodal texts usually functions as the 'guide rhythm' (van Leeuwen 1985) with which the other modes are synchronized, but polyrhythmic relations between modes are also possible. In films, speech provides the guide rhythm in dialogue scenes, action in action scenes and music may

also provide the guide rhythm ('cutting to music'). In everyday 'language in action', similarly, action provides the guide rhythm, while in 'language as reflection' speech will provide the guide rhythm. While in the second case the body movements of actors, whether in films or everyday life, will indeed be 'paralinguistic', in the first case language could be said to be 'para-actional'.

Figure 1 shows a transcription of the opening scene of *The Gruffalo* (Max Lang and Jakob Schuh 2009), an animated short film based on Julia Donaldson's famous children's book (it can be watched on dailymotion.com/video/x8tur7g). I owe the example to Thu Ngo, who has used it extensively in her work, e.g. Ngo and Unsworth forthcoming). In the transcribed segment we see the mother squirrel finding a nut. The film then cuts to her offspring – two little squirrels run, giggling, along the large branch in front of their tree home, stop for a moment as they hear a strange grunting sound, run again, and stop again as they hear the hoot of an owl (which will cause them to hurry back to the safety of their tree home, shouting 'Mama, Mama!').

The top row of the transcript indicates the squirrels' actions, the second row their 'dialogue', the third row the accompanying music, and the fourth row the sound effects. The vertical lines indicate the placement of the edits, and the fifth row indicates the shots (CS for 'close shot', MCS for 'medium close shot', MLS for 'medium long shot', LS for 'long shot', and VLS for 'very long shot').

In this sequence, music forms the guide rhythm – a soft, peacefully tinkling guitar playing phrases that repeat a three-note arpeggio four times (the rhythmic accents are indicate by a '>' above the accented notes. Each new phrase (indicated by square brackets) transposes the pitch level of this pattern upwards or downwards, so clearly demarcating the phrase boundaries. Because the musical pattern repeats itself without 'going anywhere', it creates a kind of suspense, making us wait for

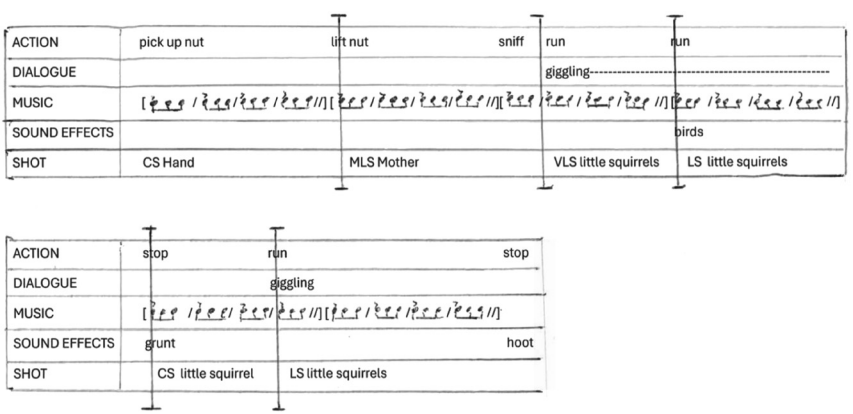

Figure 1: Rhythmic analysis of a music-based scene from *The Gruffalo* (Lang and Schuh 2009).

something to happen, something to change (indeed, when danger eventually looms for the squirrels, the music becomes more dramatic). The transcript shows how the movements of the mother coincide exactly with the beginnings of the musical phrases, though the 'sniffing phrase' is a little longer than the preceding phrases, so as to indicate the end of the 'nut gathering' move and the transition to the next move in which the squirrels will be running happily out of their tree home. Note how the edit to a close shot of a frightened little squirrel coincides with the scary 'grunt' sound effect as well as with the rhythm of the music.

In everyday social interaction people also attune to each other's rhythms, synchronizing with each other, whether in conversation, in actions like walking, or in dance and music. Studies by Erickson and others, using 16 mm film, have demonstrated how fine-grained this 'rhythmic entrainment' is. Figure 2 is a transcript of a family dinner conversation by Frederick Erickson (1982). 'G' is a guest (the researcher), 'M' the mother of the family, 'S' the son, and 'D' the daughter. They are discussing how expensive groceries have become and remark that at least the produce they grow in their own garden is free. Then they playfully imagine what it would be like if they could also grow lasagna, ravioli, milk, cheese and so on.

Transcribing the conversation as a musical score with 'bars' of equal duration and aligning simultaneous moments in the speech of different participants clearly brings out how much the participants were rhythmically attuned to each other. Observe, for example, how the *mmh*'s of the Guest and the Mother in bar 1 are exactly synchronous; how the Guest's *mm* in bar 8 coincides with the Son's '*n the*', how *napkin* and *night* in bar 3, '*things to*' and '*n the*' in bar 5, and '*-sagna*' and '*salt'n*' in bar 7 are all in exact sync, and how the Mother must double the rate of articulation of the four syllables of '*every other*' to keep in step with the Son's two syllables on '*n the*' in

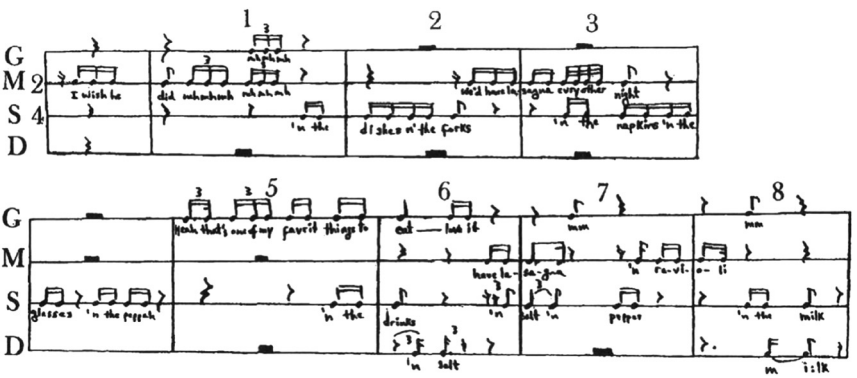

Figure 2: Rhythmic entrainment in conversation (Erickson 1982: 67).

bar 3. As Erickson (1982: 65) concluded, "Rhythm seems to be the fundamental social glue by which cohesive discourse is maintained in conversation".

Scheflen (1974: 33) has shown that rhythmic entrainment also relates to posture. During the larger sections of conversation which he calls 'positions' (e.g. explaining a point, relating a brief anecdote), a particular posture, for instance leaning forward, is held by both participants, and changed synchronously, and the same happens at the end of the position.

Finally, rhythmic entrainment has been shown to have a biological basis in neural entrainment (Lakatos et al. 2019) and seems to even extend to communication with animals (Harjunpää and Szczezepek Reed 2025).

To summarize this section, two points can be made: First, rhythm structures multimodal texts by delimiting information units and marking accents, operating at the different ranks of measure (foot), phrase, move, and phase. Second, rhythm integrates the different modes in time-based multimodal texts and the actions of the participants in social interactions and performances ('rhythmic entrainment'). For this reason, rhythm should be analysed in an embodied way, for instance by tapping on accented syllables or musical notes to 'find' the rhythm through a bodily response.

3 Sound quality, identity, and emotion

Voice quality is a semiotic resource for the expression of identity as well as for the expression of fleeting states and emotions. Identity here includes both permanent or slowly evolving identities and situationally specific roles (such as that of news-reader). As I have discussed at length elsewhere (van Leeuwen 2022) identity, today, refers not only to the identity of people (whether in terms of individual personality or in terms of the group(s) to which people belong) but also to the identity ('branding') of companies and other institutions – the London company Acapela, for instance, offers its clients synthesized voices for announcements and 'product enhancement', encouraging them to "use your own exclusive voice" and "enhance your brand" and offering them a wide range of "voices with accents, voices of celebrities, voices that surprise you with their naturalness and custom-made voices".[1]

Following van Leeuwen (1999), the parameters of voice quality are:

- Tense/lax
- Rough/smooth
- Breathy/non-breathy

1 https://www.acapela-group.com (accessed 10 August 2025).

– Soft/loud
– High/low (pitch level)
– Fast/slow
– Wavering/steady
– Nasal/non nasal

These parameters are not binary choices but scales, ranging from, e.g., maximally tense to maximally lax, or maximally fast to maximally slow, and so on. And they are *simultaneous* qualities – any voice quality has a value on each of these scales, a degree of tenseness as well as a degree of roughness as well as a degree of breathiness, and so on. Voice quality can therefore not meaningfully be represented as a system network.

Three further parameters can be added: frontal/back, high/low (articulation), and open/closed. These parameters have traditionally been seen as aspects of vowel quality. Firthian phonology extended them to the dynamic prosody of syllables (cf. Abercrombie 1965; Palmer 1970). However, they can operate as a semiotic resource at all phonological ranks. In *Star Wars – The Phantom Menace* (George Lucas 2000), for instance, the treacherous Viceroy of Naboo, a character with an inscrutable fish-like physiognomy, not only has a vaguely Chinese accent, but also speaks with a stiff jaw and almost closed mouth, so that he 'holds back' as it were, which adds to the Western stereotype of Asian inscrutability.

The meaning potential of voice quality can be based on two semiotic principles.

(1) Experiential metaphor

Here meaning is based on our physical experience of producing particular voice qualities, for instance vocal tension, and our knowledge of when these voice qualities occur, for instance, in the case of tension, when we are anxious or excited. Literal tension can then become metaphorical tension, and the context can narrow this down further. In his study of singing styles across the globe, Lomax (1968: 194) noted that in societies where women are severely repressed (very early marriages, clitoridectomy, harsh punishments for adulterous women, etc.), singing tends to be both very tense and very nasal:

> It is as if one of the assignments of the favoured singer is to act out the level of sexual tension which the customs of the society establish as normal. The content of this message may be painful and anxiety producing, but the effect upon the culture member may be stimulating, erotic and pleasurable, since the song reminds him of familiar sexual emotions and sexual emotions and experiences. (Lomax 1968: 194)

(2) Provenance

Here meaning is based on knowing where a particular configuration of voice quality values 'comes from', e.g. recognizing an iconic singing style such as Bob Dylan's or the iconic vocal style of an actress like Marilyn Monroe, or being able to 'place' a voice in terms of social class, age, gender, ethnicity, and then associate meanings with that 'place of origin'.

Hence voice quality is not a systematically organized semiotic resource. Meaning is assigned on the basis of our bodily experience as speaking and singing beings (experiential metaphor), or on the basis of our cultural knowledge (provenance) – and it is always coloured in by the situational and/or socio-cultural context.

In a study of Wallonian puppeteers, Gross (2000) asked puppeteers how they found voices for their puppets. He found that the puppeteers used voice quality to signify group identity and individual character as well as to express emotions of the moment:

> Worker and peasant puppets tend to have higher pitched voices and faster tempo than upper class characters. Upper class voices are also louder […]. (Gross 2000: 221)
> [A] self-confident character speaks in a clear voice which is not breathy […], but a character who is morally self-confident with a clear voice will adopt breathy voice when he finds himself in a subordinate position, especially when frightened. (Gross 2000: 233)
> Breathiness also shows up in people who are sad. […]. More contained sadness might manifest itself in a tremulous voice. Happy characters on the other hand, often speak through spread lips […]. The voice of someone who is emotionally tense is loud and sharp. Both of these qualities are intensified when the character is angry. (Gross 2000: 233)

Michel Chion (1999: 83) has, similarly, described how film actors use their voice to create characters:

> Most of the characters [in Bresson's *A Man Escaped*] avoid any projection of the voice and hardly move their lips – they are prisoners, under constant surveillance, not allowed to speak. On the other hand, the jailers, the masters, speak in a kind of barking, which makes the space of the prison resonate theatrically. (Chion 1999: 83)

And elsewhere (van Leeuwen 2009: 73) I have described the voice of Marlon Brando in *The Godfather* (1972) as follows:

> It is a comparatively high voice, and men tend to use high voices to dominate. It is also hoarse and rough, signalling the Godfather's harsh and unforgiving side. And it is articulated with a stiff jaw and an almost closed mouth, suggesting an unwillingness to 'give' that has us guessing what he might be keeping from us. Yet it is also soft and breathy, at times almost a whisper, making the Godfather's menacing presence disturbingly intimate and attractive. (van Leeuwen 2009: 73)

In the first decades of sound film, actors' voices have often been constant across their different roles, instantly recognizable, and expressing on screen personalities that were thought to reflect their 'real', off screen personalities. As a result, the voices of iconic actors could become models for people to emulate, so as to identify with the personalities of iconic actors, and the values they stood for, as expressed in the roles they played. Marilyn Monroe, for instance, used a high, yet breathy voice, combining feminine vulnerability and seductiveness, Lauren Bacall used a low, sensuous voice. In her autobiography she recalls how director Howard Hawks conceived of her character in *To Have and Have Not* (1944) as "a *masculine* approach – insolent. Give as good as she got. No capitulation, no helplessness" (Bacall 1979: 87). To this end Hawks not only invented 'the look' – a quizzical, look upwards with the head slightly bowed, suggesting feminine deference as well as insolence – but also told her to work on her voice: "Practice shouting, keeping the register low".

More recently, actors have begun to adjust their voices to the roles they are playing, as in the case of Brando's Godfather. As Chion (1999: 173) observed:

> Compare two roughly contemporaneous Dustin Hoffman movies. In Barry Levinson's *Rain Man*, he has a metallic and nasal voice and in Stephen Frears' *Hero* it is coarse. If you listen to both films without the picture, it is quite difficult to identify both voices as coming from the same actor. (Chion 1999: 173)

And he (Chion 1999: 174) concluded:

> The voice is ceasing to be identified with a specific face. It appears much less stable, identified. This general realization that the voice is radically other than the body that adopts it (or that it adopts) for the duration of a film seems to me to become of the most significant phenomena in the recent development of the cinema, television and audiovisual media in general. (Chion 1999: 174)

The same has happened with the singing voice. Shepherd (1991: 167), for instance, has described female singing styles in popular music, offering, as is so often the case with female 'characters', just two conflicting models. In the singing style he characterized as the style of 'woman as emotional nurturer', the voice is soft and warm, relatively low, with an open throat, and using the resonating chamber of the chest, so that the voice literally comes from the region of the heart or breast (as it also does in softer, lighter male singing styles, e.g. Paul McCartney). But in the voice of 'woman-as-sex-object' (Shepherd 1991: 167):

> [T]he softer, warmer, hollower tones of the woman singer as emotional nurturer become closed off with a certain edge, a certain vocal sheen. Tones such as produced by Shirley Bassy in 'Big Spender', for instance, are essentially head tones, and it could in this sense be argued that the transition from woman the nurturer to woman the sex object represents a shift physiologically coded from the feminine heart to the masculine head. (Shepherd 1991: 167)

But here, too, singers have come to adjust their voices to the content of the lyrics. Madonna was perhaps the first singer who used quite different voices for different songs, and even within specific songs (cf. van Leeuwen 1999: 150–152). Here is an analysis of part of 'Live to Tell' (1986):

I tell the tale to tell *Sometimes it gets so hard to hide it well* *I was not ready for the fall* *Too blind to see the writing on the wall*	lax, soft, low, smooth, steady, breathy, non-nasal
A man can tell a thousand lies *I've learnt my lesson well* *Hope I live to tell* *The secret I have learnt* *Till then it will burn inside of me*	increased tension, loudness and pitch; vibrato on some key words; non-breathy; non-nasal

Initially her voice is relatively low and relaxed, and softer, breathier and warmer than, for instance, in 'Like a Virgin'. Only after a while do an occasional vibrato and increased tension add the strength of relived emotion to some of the key moments. At this point two other (instrumental) voices join, as though representing two other women with the same or similar experiences: a tense, high voice, playing around the melody in staccato notes, and a more strident and sustained voice, doubling the melody.

Traditional TV talk show hosts such as Michael Parkinson and Michael Aspel sought to find the 'non-celebrity humanity' in the celebrities they interviewed, to bring out the supposed unity between their on-stage and off-stage personality. As Michael Aspel said in an interview (Bell and van Leeuwen 1994: 221), this is difficult with a singer like Madonna:

> Madonna would be anathema to that kind of thing. I think if we had Madonna on the show, it would be a lot more like a political interview. We would actually have every excuse to be firm and direct and challenging because, like with a politician, you don't know whether there is any truth in there at all, and so you have every right to demand to know what it is. (Bell and van Leeuwen 1994: 221)

But he was fighting a rearguard action. Today we have moved from the era of stable, enduring identities to the era of flexible and changing identities.

Finally, the sound qualities I have discussed can also be found in musical instruments and the way they are played, and in the sound effects and atmosphere tracks of films and videos. Michel Chion (1999: 84) gives an example from Bresson's *The Devil, Probably* (1977):

> At the beginning of the film, a *bateau-mouche* goes by on the Seine at night, its motor emitting a low muffled sound. We hear this deep archaic sound as the voice of the Ancestor. It evokes an African ceremonial instrument used in certain magic traditions, precisely to summon the Ancestor. This voice can also be heard in the extraordinary groans produced by trees cut down with chain saws (in the sequence of the environmental documentary), and in the rumbling of elevators. All these noises have a muted, dramatic, even deeply moving quality in the film. (Chion 1999: 84)

In conclusion, four points can be made. Sound quality is multimodal and operates at different phonological ranks, from the 'impulse' (e.g. syllable or musical note) upwards to the 'phase' and beyond. It can express more or less permanent or slowly evolving identities as well as situationally specific roles and more or less momentary states or emotions. As a semiotic resource, it is not (yet) systematically organized. Instead, its motivated meanings are understood on the basis of our physiological experience as speaking and singing beings, and on the basis of cultural knowledge. And finally, it is an embodied form of meaning-making and should therefore be analysed with the body, by feeling it, so as to represent how it is perceived by human beings rather than by machines.

4 Pitch contours, language, and music

Derycke Cooke (1959) in his pioneering book *The Language of Music* discussed the melodic structure of musical phrases and distinguished 16 basic pitch contours. He documented the extraordinarily consistent and continuing use of these contours over five centuries, in classical as well as in (Western) popular music – and also found considerable consistency in the kinds of lyrics that went with these contours in vocal music. Figure 3 shows a few examples of one of his contours, the 1-4-5-6 contour – the numbers refer to the intervals between the rhythmic accents. The 1-4-5-6 contour therefore begins with a fairly large rise, followed by to smaller rises.

As in language, melody is therefore based on the fundamental opposition between up and down, rise and fall, ascending and descending. According to Cooke (1959), ascending melodies express 'outgoing', 'active' emotions and descending melodies more 'incoming', 'passive' emotions, while 'arched' melodies combine the two. He grounds this in bodily experience. Ascending in pitch requires an increase in vocal effort, descending in pitch allows us to decrease vocal effort. Both can involve large or small intervals, which is, again, grounded in bodily experience – increasing pitch with small steps requires less vocal effort than increasing it by large steps.

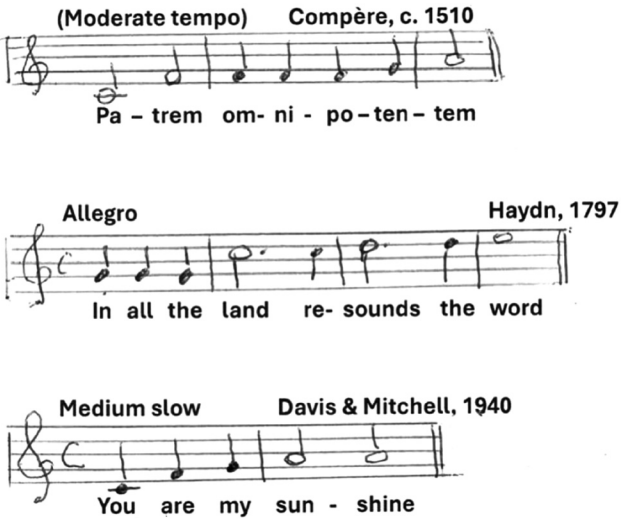

Figure 3: 1-4-5-6 musical pitch contour (Cooke 1959).

It is no different in language. As Bolinger (1986: 194) expressed it: "intonational melodies have their symbolizing power thanks to a primitive drive mechanism that raises pitch as tension rises and lowers it as tension falls", and its "primitive and still surviving function is the signalling of emotion" (Bolinger 1986: 195). Therefore, "the grammatical functions of intonation are secondary to the emotional ones; speakers feel differently about what they say, and the feelings manifest themselves in pitch changes" (Bolinger 1986: 27).

Halliday (1970: 6) did recognize the "attitudinal" potential of intonation: "It is helpful to think of attitudes and emotions as part of meaning, to consider that all intonation patterns convey meaning and then to ask what kind of meaning they convey". But the information structuring function of intonation is central in his intonation theory, and attitudinal meanings are said to be expressed by 'secondary tones': "intensified forms (which) indicate a greater intrusion on the part of the speaker, by which he is expressing either an attitude (surprise, indignation, unconcern, sarcasm, etc.) or some definite connection between what he is saying and something else in the discourse (contrast, contradiction, unpredictability, etc.) or some combination of the two" (Halliday 1970: 24).

A multimodal approach to pitch contours, however, must focus on what is common between different modes, which, in the case of pitch contours is the

expression of emotion, and its grounding in bodily experience. Even in language, intonation "is fundamentally the opposition of up and down, with meanings clustering around the poles of the opposition, in accord with metaphorical extensions" (Bolinger 1986: 221).

In both music and language, pitch movement can express momentary emotions as well as styles that characterize particular genres, the habitus of social groups, or the key values of whole cultures. As Lomax (1968: 136) writes in his cross-cultural study of singing styles:

> The use of wide intervals may symbolize a less confining, freer, more wide-ranging approach to the use of space (social and/or ecological) in a society where access to life resources is open to all members of the community on more or less equal terms (while) prominence of very narrow intervals turns up in cultures whose members are confined spatially or restricted by a system of rigid status differences in the use of productive and social resources. (Lomax 1968: 136)

The examples in Figure 4 show how the melodies of language and music may express similar meanings in similar ways – the notation of the speech patterns is adapted from Delattre (1972) and uses musical staves to give a sense of the scale of the pitch steps rather than indicate precise pitches (the examples should be read in an embodied way, by speaking or singing them, 'feeling' them). The first example compares a music announcement from a top 40 commercial music radio station with an 18th century hymn written by Charles Wesley. Both seek to energize and enthuse the listener, to enthuse them, in the one case, for the song, in the second case originally for the Methodists' fight against their critics, later perhaps for missionary endeavours. Each part of the message steps up with big leaps. In the case of the announcement the steps are 'lot of money', 'America', and 'really big name', in the case of the hymn, "soldiers of Christ", "arise", and "put your armour on".

The second example is a music announcement from an 'easy listening' music radio station. Here the key moments ("watch", "happens", "too") stay at the same relatively low level, and only the first part of the performer's name, "Lucio" rises slightly, after which the melody descends. The second example is from a 19th century hymn written by the English poet Frances Havergal, where the lyrics suggest a kind of surrender of the self to God. Its two parts (the first starting with "take my life", the second with "consecrated") gradually descend with small steps.

Fónagy and Magdics (1972: 304) found, not only that "emotions are expressed in European vocal and instrumental music by a melody configuration, dynamics and rhythm similar to those of speech" but also found very similar pattens in data from two quite dissimilar languages, Hungarian and French, and concluded that these patterns were not arbitrary and not language-specific: "If a certain emotion is expressed by similar melodic patterns in non-related languages, their intonation must not be considered arbitrary" (Fónagy and Magdics 1972: 292). They also stressed

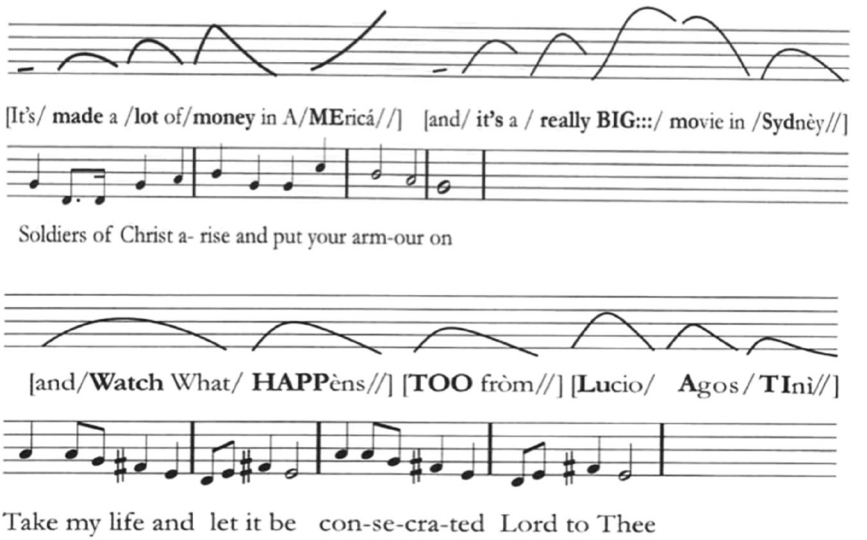

Figure 4: Ascending and descending pitch patterns in speech and music.

that the expression of emotion results, not from melody alone, but from configurations of melody, tempo and sound quality, whether in speech or in music. In Figure 5, I give examples of their description of the expression of 'joy' and 'tenderness'. 'Joy' they described as characterized by a wide pitch range at a high pitch level. The melody rises, then falls sharply, then stays level or descends slightly. The tempo is lively. 'Tenderness' they described as characterized by a high pitch level but with a narrow pitch range, while the melody descends slightly and undulates. The tempo is medium. In singing a soft, slightly nasal and labialized voice is used, in instrumental music an instrument with a similar quality. The first musical example is a phrase from 'Joy Spring' by the jazz trumpetist Clifford Brown, the second example is from 'Tenderly' by Walter Gross and Jack Lawrence.

Figure 6 gives two further examples, 'surprise' and 'anguish'. According to Fónagy and Magdics (1972), 'surprise' is characterized by a sudden upwards (or up and down) glide to a high pitch level, followed by a fall. The extent of the fall then depends on the degree of the surprise. The tempo is medium, the voice breathy. 'Anguish' they describe as characterized by an extremely narrow pitch range at mid pitch level. The melody rises about a semitone and then returns to mid-high level "where it becomes so to speak paralysed" (Fónagy and Magdics 1972: 289). The voice is tense and breathy. The first musical example is a phrase from 'Sudden Samba' by the American jazz pianist and composer Neil Larsen, the second is a phrase from Ray Charles' 'My Heart Cries for You'.

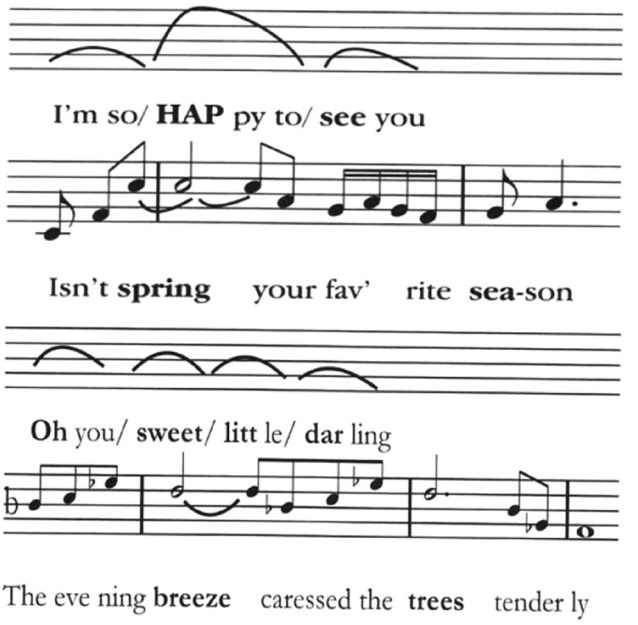

Figure 5: 'Joy' and 'tenderness' in speech and music.

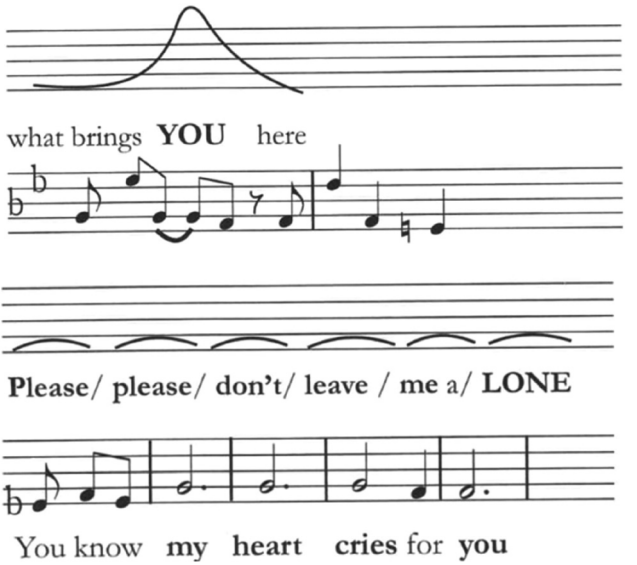

Figure 6: 'Surprise' and 'anguish' in speech and music.

In these examples, the titles and/or lyrics express the same emotion as the melodies. But this is not always so. It is quite possible to say 'I'm so happy' while sounding decidedly unhappy, and in many songs the emotions expressed by the lyrics contrast with those expressed by the music. In van Leeuwen (1991), I analysed Simon and Garfunkel's 1965 hit song 'I am a Rock' in this vein. While the lyrics start by expressing a wistful sadness ("A winter's day/In a deep and dark December/I am alone") and then move to a refusal of love, based on the fear of being hurt ("I have no need of friendship/Friendship causes pain"), the music has a bright rhythm, and a happily ascending melody, free of any sentimental devices, making the refusal of love defiant rather than sad.

5 Conclusions

To conclude, an embodied, multimodal phonology does away with the split between articulatory phonetics and semantics, the materiality of sound and its meanings, and studies the meaning potentials of sound on the basis of their physiological and physical articulation in whatever medium of expression. In addition, it is always multimodal, embracing all sound media and looking for principles that can be applied to all, as I have tried to do throughout this article. It also sees 'expression' as not only realizing lexicogrammatical structures, but as itself meaningful, adding a layer of different meanings, with sound quality and pitch contours contributing to the realization of identity meanings and emotions. And it sees rhythm, rather than intonation, as realizing information structure in speech, and indeed also in all other time-based modes.

To a large degree, these ways of making meaning reflect emerging patterns of social communication, which therefore require a new approach to semiotics in which, as Gunther Kress (2011: 55) formulated it, "material signifiers carry a set of affordances from which sign-makers and interpreters select according to the communicative needs and interests in a given context". This aptly describes, for instance, how meaning is made in contemporary identity design which, like consumer goods, must be customized and individualized, and always aim for novelty and innovation. The new 'phonology' outlined in this paper is just one of the semiotic resources whose meaning-making reach has extended to meet this demand. But this 'creative' approach to meaning making must still be carried by a functional skeleton. Just as today's buildings are functionally very similar, often assembling ready-made modules, but do so behind ever different facades, whether the retained facades of historic buildings or decorative screen claddings, so the functional structures of texts and other semiotic artefacts are becoming increasingly generic and similar, while their graphic and/or phonic design becomes increasingly diverse so as to enable the

expression of increasingly many different identities. Contemporary semiotic resources such as PowerPoint demonstrate how functionality and identity work together – embellishing the same fundamental 'headline plus bullet points' format with designs that allow the expression of a wide range of values – 'modern' and 'future-oriented' or 'traditional' and 'retro', 'business-like' and 'technical' or 'organic' and 'eco-conscious', 'concise' and 'focused' or 'elaborated with flourishes' and 'aesthetic', and so on – all values, of course, that matter in the global corporate culture which has brought PowerPoint into the world.

Research ethics: Not applicable.
Informed consent: Not applicable.
Conflict of interest: The author declares that there is no conflict of interest.
Data availability: Not applicable.

References

Abercrombie, David. 1965. *Studies in phonetics and linguistics*. London: Oxford University Press.
Abrams, Kenneth & Thomas G. Bever. 1969. Syntactic structure modifies attention during speech perception and recognition. *Quarterly Journal of Experimental Psychology* 21(3). 28–29.
Bacall, Lauren. 1979. *Lauren bacall by myself*. New York: Random House.
Bell, Philip & Theo van Leeuwen. 1994. *The media interview: Confession, contest, conversation*. Sydney: University of New South Wales Press.
Bolinger, Dwight W. 1986. *Intonation and its parts: Melody in spoken English*. Stanford: Stanford University Press.
Brown, Gillian. 1977. *Listening to spoken English*, 2nd edn. London: Longman.
Chion, Michel. 1999. *The voice in cinema*. New York: Columbia University Press.
Cooke, Deryck. 1959. *The language of music*. London: Oxford University Press.
Couper-Kühlen, Elizabeth. 1993. *English speech rhythms: Form and function in everyday verbal interaction*. Amsterdam: John Benjamins.
Daneš, František. 1960. Sentence intonation from a functional point of view. *WORD* 16(1). 34–54.
Delattre, Pierre. 1972. The distinctive function of intonation. In Dwight L. Bolinger (ed.), *Intonation: Selected readings*, 159–175. Harmondsworth: Penguin.
Erickson, Frederick. 1982. Money tree, lasagna bush, salt and pepper: Social construction of topical cohesion among Italian Americans. In Deborah Tannen (ed.), *Analysing discourse: Text and talk*, 43–70. Washington, DC: Georgetown University Press.
Fónagy, Ivan & Klara Magdics. 1972. Emotional patterns in intonation and music. In Dwight L. Bolinger (ed.), *Intonation: Selected readings*, 288–305. Harmondsworth: Penguin.
Gross, Joan. 2000. *Speaking in other voices: An ethnography of Walloon puppet theaters*. Amsterdam: John Benjamins.
Halliday, Michael A. K. 1970. *A course in spoken English: Intonation*. London: Oxford University Press.
Halliday, Michael A. K. 1985. *Spoken and written language*. Geelong: Deakin University Press.
Harjunpää, Katariina & Beatrice Szczepek Reed. 2025. Prosodic matching beyond humans: On the interactional basis of "cat-directed" talk. *Language & Communication* 103. 65–85.

Kress, Gunther. 2011. *Multimodality: A social semiotic approach to contemporary communication*. London: Routledge.
Lakatos, Peter, Joachim Gross & Thut Gregor. 2019. A new unifying account of the roles of neuronal entrainment. *Current Biology Review* 29(18). 890–905.
Lehiste, Ilse. 1977. Isochrony reconsidered. *Journal of Phonetics* 5(3). 253–263.
Lenneberg, Eric H. 1967. *Biological foundations of language*. New York: John Wiley & Sons.
Lomax, Alan. 1968. *Folk song style and culture*. Washington, DC: American Association for the Advancement of Science.
Martinec, Radan. 2000. Rhythm in multimodal texts. *Leonardo* 33(4). 289–297.
Martinec, Radan. 2002. Rhythmic hierarchy in monologue and dialogue. *Functions of Language* 9(1). 39–59.
Meyer-Eppler, Werner. 1957. Realization of prosodic features in whispered speech. *Journal of the Acoustical Society of America* 29(1). 104–106.
Ngo, Thu & Len Unsworth. forthcoming. *Digital multimodal adaptations of children's literature: Semiotic analyses and classroom applications*. London: Routledge.
Palmer, Frank R. 1970. *Prosodic analysis*. London: Oxford University Press.
Pearlman, Karen. 2009. *Cutting rhythms: Shaping the film edit*. London: Focal Press.
Pike, Kenneth L. 1945. *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.
Pike, Kenneth L. 1962. Practical phonetics of rhythm waves. *Phonetica* 8(1–3). 9–30.
Scheflen, Albert E. 1974. *How behavior means*. New York: Jason Aronson.
Shepherd, John. 1991. *Music as social text*. Cambridge: Polity Press.
van Leeuwen, Theo. 1982. *Professional speech: Accentual and junctural style in radio announcing*. Sydney: Macquarie University MA thesis.
van Leeuwen, Theo. 1985. Rhythmic structure of the film text. In Teun A. van Dijk (ed.), *Discourse and communication: New approaches to the analysis of mass media discourse and communication*, 216–232. Berlin: De Gruyter.
van Leeuwen, Theo. 1991. The sociosemiotics of easy listening music. *Social Semiotics* 1(1). 67–80.
van Leeuwen, Theo. 1992. Rhythm and social context: Accent and juncture in the speech of professional radio announcers. In Paul Tench (ed.), *Systemic phonology*, 231–262. London: Pinter.
van Leeuwen, Theo. 1999. *Speech, music, sound*. London: Palgrave Macmillan.
van Leeuwen, Theo. 2005. *Introducing social semiotics*. London: Routledge.
van Leeuwen, Theo. 2009. Parametric systems: The case of voice quality. In Carey Jewitt (ed.), *The Routledge handbook of multimodal analysis*, 68–77. London: Routledge.
van Leeuwen, Theo. 2022. *Multimodality and identity*. London: Routledge.
van Leeuwen, Theo. 2025. *Multimodality and time: A social semiotic approach*. London: Routledge.