Logopedics Phoniatrics Vocology. 2005; 30: 100–107

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Does the acoustic waveform mirror the voice?

STEN TERNSTRÖM

Department of Speech, Music and Hearing, School of Computer Science and Communication, Kungliga Tekniska Högskolan, Stockholm, Sweden

Abstract
Over recent decades, much effort has been invested in the search for acoustic correlates of vocal function and dysfunction. The convenience of non-invasive voice measurements has been a major incentive for this effort. The acoustic signal is a rich but also very diversified source of information. Computer literacy and technical curiosity in the voice care and voice performance communities are now higher than ever, and tools for voice analysis are proliferating. On such a busy scene, a review may be useful of some basic principles for what we can and cannot hope to determine from non-invasive acoustic analysis. One way of doing this is to consider communication by voice as though it were engineered, with layered protocols. This results in a scheme for systematizing the many sources of variation that are present in the acoustic signal, that can complement other strategies for extracting information.

Key words: *Channel code, speech communication, transport protocol, vocology, voice analysis*

Introduction

The established methods of scientific inquiry rely on dividing reality into slices that are small enough to grasp and to study. The resulting insights depend on the direction of the cut. Various aspects of communication by voice have been studied by many disciplines, and several schemes for slicing its reality have been proposed. Yet another is presented here. Most of the components of this scheme are borrowed from communications engineering and speech communication research; but perhaps some of its combinations are original, so that the reader herein may find alternative ways of thinking about some familiar problem in voice research.

The sounds of speech or song carry a wealth of information. Even when we cannot see the person we are listening to, we quickly arrive at some idea of her age, gender, nationality, dialect, personality, mood, educational and social background, and so on; and often these judgments will be accurate. As health or voice professionals, we may also be able to judge certain aspects of the speaker's health in general, and the condition of her voice in particular. Superimposed on the spoken words are dozens of cues which lead us to infer a context, right or wrong. This context, set against our prior experience, constrains and assists us in interpreting what we hear.

For decades now, it has been said that non-invasive voice analysis based on the acoustic signal should have great potential for being a convenient and inexpensive tool for investigation. The multitude of things that we believe that we can hear in a voice seems to point to this conclusion, and much progress has been made; but our machines are still far from being listeners. Of all the information that is represented in the sound, some is relevant to voice function, but more of it is not; and that which is, can only be interpreted when a correct and sufficiently complete set of non-acoustic background information is available. What kinds of information might be extractable from the acoustic voice signal, in principle? How far can we unravel the signal's properties, if we try to engineer the voice communication process in reverse?

The concept of layered protocols

Most people use their voice primarily for communicating with others. 'Communication' is a broad

Correspondence: Sten Ternström, Kungliga Tekniska Högskolan, School of Computer Science and Communication, Department of Speech, Music and Hearing, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden.

(accepted 1 July 2005)

ISSN 1401-5439 print/ISSN 1651-2022 online
DOI: 10.1080/14015430500238400

term. Here, we will use it to refer generally to the process of transferring messages, of any kind at all, from one party to another. Engineers who design reliable systems for communication must rely on very structured approaches. Suppose that we wish to communicate a text message to a receiver across some physical distance. Here is how an engineer might chart the procedure. The example is historical: manual telegraphy, in Morse code.

- A sender wishes to send a message to transfer certain ideas or information.
  - ↓ The sender uses a *language* to cast the message into a *script* with an agreed syntax (for example, English). The script contains a sequence of *symbols* (letters).
    - ↓ The symbols are translated into *channel code* (dots and dashes).
      - ↓ A *transducer* turns the channel code into *physical events* (short and long beeps) in a *medium* (such as a cable, or radio waves).
        - → The physical events propagate through the medium to the *receiver*.
      - ↑ Another transducer, at the receiver, recovers the channel code from the physical medium (dots and dashes).
    - ↑ The channel code is reassembled into the sequence of symbols, i.e., the script (letters).
  - ↑ The receiver interprets the script according to the syntax of the agreed language.
- The receiver understands the message.

In transmission, the message passes down through several layers of representation, into a physical medium, and then up again at the receiver, through the same layers, but in the reverse order. Notice that if we substitute the physical layer with flashes of light instead of beeping tones, the transferred message remains the same. Or, the language can be French instead of English, which is of no consequence for the physical layer. This same approach is also used for *storage* of information, in which case the physical transitions are made to persist in the medium, rather than to propagate in it.

In modern technical communication systems, such as computer networks, there are many more layers, or *levels of protocol*, than in this example. Additional layers can provide useful functions such as error correction, encryption and routing. Each layer must know exactly *what* to exchange with the next layer, but it must know nothing of *how* its neighbouring layers do what they do. The machinery at each level must communicate only through carefully defined interfaces, which are laboriously spelt out in every detail by standards committees. If this were not so, it would not be possible to improve or modify one layer at a time, without impacting the others.

Layered protocols in speech

Human communication by sound, too, can be thought of as if it were built on a layered transport protocol; indeed, this is reflected by the set of disciplines that attend to speech. For example, we could say that the message originates in a thought (psychology), the message script is composed of words (linguistics), the symbols are the phonemes (phonology), the channel code corresponds to the phonemes (speech perception/phonetics), the sender's transducer is the voice organ (vocology), and the medium is the air in which sound propagates (acoustics). In human communication, however, the protocols are established not by committees, but by everyday conventions; and they are acquired mostly through learning by experience. Another difference is that, unlike a man-made system, our brain has simultaneous access to nearly all the levels in this communication. Our appreciation of a given song can be heightened by superior acoustics, a skilled singer, a talented lyricist, an inspired composer, healthy ears, a fine hi-fi system, and an extensive experience as listeners. As listeners, we can also choose to ignore obvious shortcomings of the transport layers.

As an exercise, let us try to define in some detail the 'protocol layers' in spoken communication, as though speech were an engineered procedure. One tentative scheme is suggested in Figure 1. The many readers who are more knowledgeable than the author in the traversed disciplines will no doubt want to rearrange it and improve on it, and alternative schemes have been suggested by others (1). Nevertheless, the figure will serve as a roadmap for the rest of this section. Protocol levels will be referred to in parentheses.

Before we deal with the message, we need somehow to account for the influence of overall context, as well as of speaker characteristics. For convenience, we assume that the communication chain does not start with the message, but rather with a context (protocol level 1), in which the sender (level 2) lives and acts and finds reason to create messages (level 3).

Unlike text messages, a spoken message typically contains several types of information, which are sent in parallel: the words, the prosody, non-linguistic sounds, and extra-auditory information, such as gestures and facial expressions (2). For our analogy

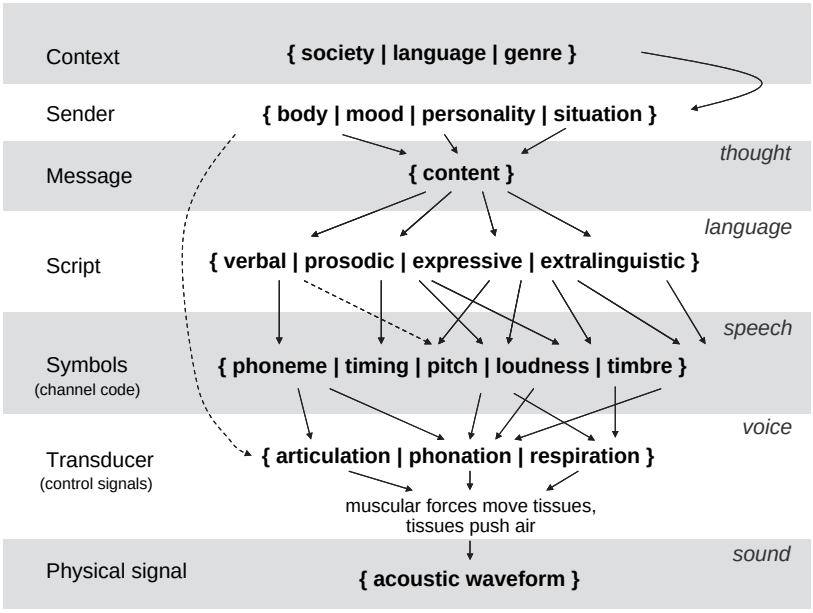

Figure 1. A suggested division of sender-side protocol layers in spoken communication.

with a technical system, the different aspects of the message can be thought of as being spread out across *multiple streams* of information, each with its own script (level 4). For example, the prosody 'script' would consist of contours for emphasis and inflection, generated by intonation rules of the language and by its rules for emphasis. Both auditory and extra-auditory streams can be used to transmit supplementary encodings of similar information, as when we reinforce spoken language with body language. The number of available streams will depend on the situation, for example, on whether or not the sender and receiver can see each other.

The auditory parts of the message are to be perceived by hearing, so the encoding of the auditory streams must at some level be perceptual. Auditory perception is in itself probably a multi-layered process, many details of which are not known, so the delineation of protocol layers on the receiver side becomes a bit speculative. We do know that the perceptual dimensions *loudness*, *pitch* and *timbre* are relevant to speech, that *phonemes* can be regarded as perceptual units, and that *timing* is an important component of how prosody is perceived. The listener must perceive the intended sequence of phonemes, each presented at a controlled time, pitch, loudness and timbre.[1] A perceived deviation in any one or more of these properties may change the meaning of the message, and each property can assume a range of values fairly independently of the others.

In principle, then, these five time-varying properties—phoneme, timing, pitch, loudness, and timbre—are perceptual variables that can function as separate communication channels. In practice, there will be a lot of covariation in the information in these channels. For example, pitch and loudness are both strong correlates of prosodic emphasis, because both respond to subglottal pressure variation. The engineer would note that such covariation in the channels provides redundancy, making the speech signal more robust.

So, the message is distributed over several streams of information, which *in turn* are encoded simultaneously into several perceptual variables. This means that the values of each of these variables result from a superposition of data from the different streams. As if that were not enough, the variables also embed underlying contextual information, encoded at different time scales. A deep voice has a low habitual pitch (low average F0) and a dark timbre (the upper formant frequencies are low), both of which are cues that imply a large person, usually a male. The habitual pitch is modulated by the prosody of the utterance. The pitch, at any given moment, thus provides a combination of contextual information and message information. Such 'overloading' is present not only in pitch, but in the other perceptual variables as well. Typically, a baseline value or a slowly changing trend provides information about the sender and/or the context, while the message information is carried by deviations made from the baseline. Using a carrier/modulator metaphor from analogue telecommunications, this view was presented in Traunmüller's Modulation Theory (1,3).

What types of information do the perceptual variables represent? Loudness and pitch are usually

thought of as values on a *continuous* scale. In a given language, there will be a limited number of phonemes to choose from, so the phoneme inventory is *categorized* or *discrete*.[2] Timing is implemented on a continuous scale, but may be perceived categorically, for example, when duration is used to discriminate stressed syllables from unstressed. As for timbre, for the sake of this discussion, we may consider it as a catch-all holder for any remaining auditory dimensions that are not already allocated to the discrimination of phonemes. A timbre value is then a point in some abstract space, the structure of which we shall ignore; although timbre of course is highly relevant to acoustic voice analysis.

Phonemes are easy to equate with *symbols* (level 5), but what of the continuous variables? If we accept the notion that specific meanings can be encoded by temporal gestures in the continuous variables, i.e., by the *modulations*, then the transmitted data in each perceptual channel can be viewed as a sequence of symbols.

The voice organ—a complex transducer

Again comparing to Morse code, we have just discussed the layer or protocol level that corresponds to the dots and dashes. The next level down is the *transducer* (level 6), that is, the voice organ. The role of the transducer is to convert the channel code (the perceptual variables) into a sequence of physical events (the acoustic waveform), so that it can propagate in the medium to the receiver.

On the face of it, the conversion scheme is fairly straightforward. By and large, timing maps to scheduling, pitch maps to fundamental frequency, loudness maps to amplitude, and timbre maps to various spectral features. Phonemes map to predetermined configurations of phonation and articulation, such as voicing, frication, constriction, occlusion, rounding and so on. Each of these configurations leaves on the sound spectrum a specific imprint, which can be resolved by the receiver's transducer. Remarkably, all the perceptual channels exploit spectral features that spread over substantial parts of the speech spectrum, but in different ways. This is probably another reason why speech signals are robust, and difficult to distort beyond recognition.

However, while the mapping of perceptual variables to the acoustic spectrum may be fairly straightforward to describe, the machinery that makes it happen is not. Respiration, phonation and articulation are performed by many dozens of muscles that are controlled with precise timing and coordination. In learning to speak, the speaker has internalized motor programmes or 'subroutines' that perform these tasks automatically for an everyday vocabulary. This is one level where the brain, mercifully, does keep the details hidden from our conscious attention. Yet, the more detailed we can make our models of the internal workings of this transducer, the more we are likely to understand from the acoustic signal. This is a topic central to vocology.

The acoustic signal

Finally, we arrive at the physical level, the acoustic signal (level 7). The sound waves are of course likely to change on their way through various room acoustics, noises and media, but the technicalities of faithful transmission are beyond the scope of this presentation. Let us simply assume that the signal arrives at the receiver unchanged.

The voice signal in its acoustic form, as the engineer has now argued, contains several streams of information pertaining to the same message, and these streams are encoded in a spread-spectrum, highly redundant fashion that practically ensures delivery in a wide variety of circumstances. Still, from a mathematical standpoint, the acoustic signal is but a *single value* that changes rapidly with time.[3] This value, at any given point in time, represents the pressure deviation from the static atmospheric pressure. The instantaneous extent of the pressure deviation, as taken by a microphone at any particular moment, is utterly meaningless; it is the pressure *waveform*, as a function of time, that carries information. Anything and everything that we can measure from the acoustic signal will be derived in some way from this two-dimensional entity: how the instantaneous pressure changes over time.

By measuring the waveform, and subjecting it to some mathematical transforms, we may obtain basic quantitative descriptors such as the average magnitude, the periodicity, and the spectrum of the signal. These three are of particular interest, because their relationships to the percepts of loudness, pitch and timbre are fairly well known. Consequently, instruments are available that will deliver these transforms: level meters, F0 extractors, spectrographs. But the instruments provide no clues as to what the numbers *mean*, because they are designed to describe the physical level only. Comparing again to Morse code, the acoustic instruments can tell us how strong the beeps are, or what frequency they have; but this information is quite irrelevant for interpreting the symbols (level 5) or the message (level 3). For more advanced processing to be explanatory, we need detailed models not only of the transducer, but of the layers above it as well. These models will belong to various domains, including physics (level 6), percep-

tion (level 5), speech recognition (level 4), cognition (level 3), psychology (level 2) and anthropology (level 1). At each step, we will encounter problems of disambiguation: that the output we wish to analyse might have been produced in one of several ways, or for one of several reasons. Without access to a more complete context of non-acoustic data, the stairway up to the correct explanations will at some point be closed.

Fortunately for voice research, the acoustic information *is* relevant when we want to assess the quality and the operation of the *transducer*, the voice. For example, weak high-frequency partials combined with a substantial noise component may indicate a glottal insufficiency. However, because of the many layers of content that are embedded in the voice signal, we must inhibit the message traffic, or smooth it out, before we can deduce things about the transducer or the sender. That is why we so often resort to measuring a sustained vowel, which pauses the sequence of symbols (level 5); or to taking the long-time average spectrum, or plotting the phonetogram, both of which can be used so as to average out the message content (level 3), thereby to reveal some property of the sender (level 2).

Another way of putting this is to say that we can compute the statistical distributions of the low-level data at *different time scales*, thereby to obtain different views of the acoustic signal. For instance, the low-level data of SPL, F0 and spectrum can be smoothed or time-averaged as they are; or they can be subjected to further transforms at longer time scales, as when we calculate the frequency of a vibrato or tremor, or some characteristic of the long-time average spectrum (LTAS). We observed on the sender side that there is at least some correspondence between the protocol level and the time scale on which the encoding at that level occurs. In Figure 2, several acoustics-based measures are arranged on a scale of temporal magnitudes. The layout gives some idea of how the measures are potentially relevant to different aspects of vocal function or to different levels of the communication protocol.

In research on speech recognition and natural language understanding, artificial intelligence techniques have been applied in attempts to re-create a context from the signal. For example, the brain's access to multiple protocol layers has been modelled using a technique known as 'blackboarding' (4). This means that software processes at each level of protocol maintain a log of their activities on a metaphorical blackboard that is visible to the other levels. Any process is free to exploit any blackboarded information that it finds useful. While this technique is conceptually powerful, it is also computationally inefficient. Another approach is to combine pattern recognition with automatic structuring of incoming signal data into *events* (5); the events are then automatically arranged into some hierarchy that corresponds in some way to the communication levels. This is intuitively similar to

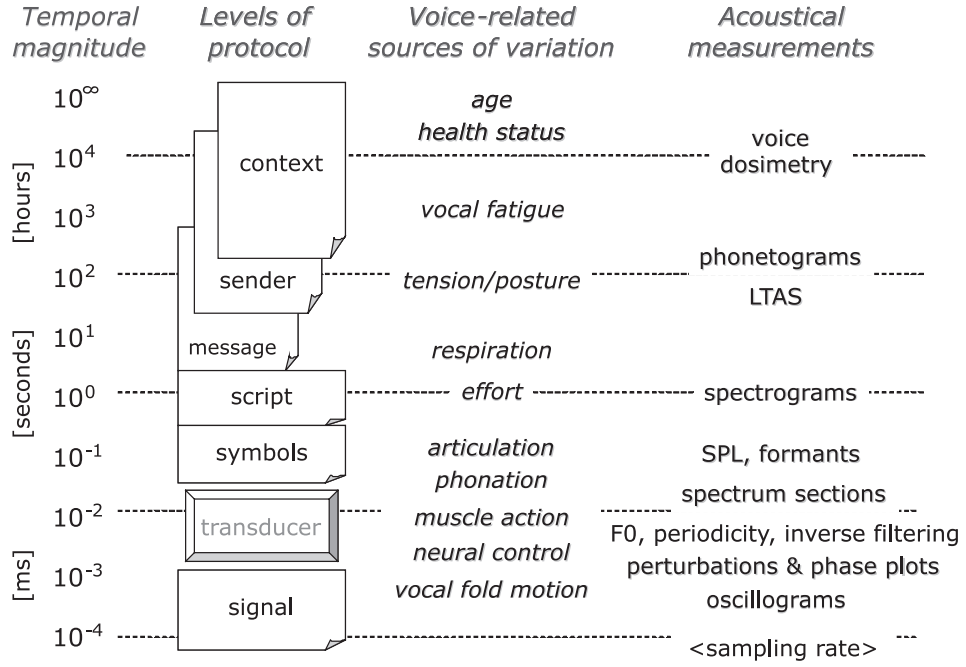

Figure 2. Approximate time scales of mechanisms of voice communication, production and analysis.

how perception works. It seems plausible that such methods could be particularly effective if they were limited to voice analysis rather than to full-scale natural language understanding. For example, a system that locates and counts the occurrences of, say, diplophonic phonation in a recorded utterance would be far more useful than one that computes a total average of a perturbation measure taken over that utterance. Such a function could be implemented by making minor modifications to current speech recognition systems.

Acoustical measures of voice

Strictly speaking, there is no such thing as an 'acoustical voice measure'. There are acoustical measures of physical phenomena, and there are ways in which those phenomena relate to certain aspects of voice function. By way of example, we will look at just two of the acoustical measures most commonly used for voice analysis, sound pressure level and fundamental frequency, and discuss how observations of their unprocessed values can relate to voice function.

Sound pressure level

The main significance of the sound pressure level[4] (SPL) is the *distance* between the sound source and the point of measurement. Sounds become weaker with increasing distance to the source, much as objects receding in our field of view become smaller to the eye. In a free field, the direct sound loses 6 dB in level with every doubling of the distance. This is why the measurement distance must always be specified when reporting the SPL. It also explains why speaking close to someone else's ear can be dangerous to their hearing. Tests have shown that listeners in free-field conditions perceive the sound pressure level mainly as a correlate of the distance to the speaker (6), while the vocal effort that is perceived by the listener is determined mainly by spectral factors.

That being said, a normally functioning voice should be able to produce at least some range of sound pressures, from soft to loud voice. If we have access to normative data for comparison, a restricted range in SPL can be evidence of a vocal problem. For example, a low maximum SPL may be a consequence of insufficient subglottal pressure, incomplete glottal adduction, glottal closure that is too gradual, or a vibrational amplitude of the folds that is somehow restricted. Each of these explanations would point to a different set of possible diagnoses. A high minimum SPL, on the other hand, would indicate a problem in getting the vocal fold oscillations started. The reasons for this could include hypertension, or vocal fold scarring, or lack of adduction. To resolve the issue, other, non-acoustic, assessments are needed.

In the acoustic and aero-acoustic domain, the sound pressure that a voice produces depends on (i) the maximum declination rate of glottal flow, MFDR; (ii) the fundamental frequency, F0; (iii) the formant frequencies, especially F1.

That is as far as the acoustic measures will take us. To explain the observed values of these measures, we must examine the non-acoustic circumstances that drive the glottal flow, determine F0, and control the formant frequencies.

The technical issues when measuring the SPL of voices include

- the total channel gain, including the microphone distance, to be calibrated for
- the frequency weighting: linear or A or C scale
- the integration time for the level computation ('slow' or 'fast' on the meter)
- averaging the level, or taking the level of the equivalent average power (L_{eq})
- averaging with or without pauses and/or unvoiced segments
- using the RMS level or (rarely) peak level

The upper SPL limit of a voice is very individual (7), and this limit becomes particularly important when it is related to the ambient noise level. The decibel scale is probably misleading when it comes to risk assessment. If, in a noisy environment, a modest increase of 3 dB in vocal output would be required for making the speech intelligible, the acoustic power radiated by the speaker would have to be *doubled*. The corresponding increases in the risk for various types of tissue damage are not known; however, the risks are more likely to be proportional to parameters of vocal fold tissue mechanics, linear or squared, than to the sound level in decibels. It seems clear that even a small reduction in ambient noise, and thus in vocal effort, should be of great help in a noisy workplace. Depending on one's profession, a weak voice may or may not be of great consequence.

Fundamental frequency

Most fundamental frequency measures reflect the short-term average of the inverse of the period time of several consecutive glottal cycles.[5] For signals from normal phonation, this measure is relatively easy to obtain, and it is rather insensitive to transmission limitations such as bandwidth or harmonic distortion. Just as for the sound level, we would expect a healthy voice to have some minimum range in fundamental frequency (F0). Abnormally high or low habitual F0, or constrained F0 limits, all may be indicative of voice problems—or of some-

thing else. Perhaps more than any other parameter, F0 is recruited for communication at several levels simultaneously: habitual pitch, prosody at the phrase and syllable levels, and emotional expression. Therefore, F0 can be biased by environmental and psychological factors.

F0 measurements are often used as a starting point for analyses of perturbations and tremor (8), which require higher precision in time than F0 in itself. The airborne signal is not ideal for determining the exact timing of each glottal cycle, especially if it contains background noise of any kind. Non-acoustical methods such as electroglottography may be preferable for this purpose. Many approaches have been tried for the analysis of period-time perturbations, but, to the author's knowledge, reliable perturbation signatures of specific voice pathologies have yet to be identified.

The phonetogram

Combining SPL and F0, we obtain the phonetogram, which shows a voice's range in both these dimensions. The automatic computerized phonetogram shows their statistical distributions as well, and can be made to colour-code any third parameter, related for instance to voice quality. In the author's opinion, the latter is a superb mapping device which deserves more widespread use, and which can be used in many more ways than we have seen so far. By resolving the combinations of F0 and SPL at which a given voice characteristic tends to change, the phonetogram creates a context of sorts, and it sorts information in a way that is clearly relevant to voice function. In addition, when used as a feedback device, the computerized phonetograph tends to encourage the user to exercise the voice, giving it a very interesting pedagogical and therapeutical potential (9).

The phonetogram does have some drawbacks. It takes a long time to acquire properly, there is a lack of standards for how to instruct the subject, and there is a risk of overdriving the subject's voice in striving to find the upper boundaries.

Conclusion

Using acoustic measures to analyse voice function requires understanding of the acoustic signal and how it is structured, and much work in this area remains to be done. Some readers may find it helpful to consider the structure that has been suggested in this paper, if only to contrast it with their own view. The important thing is to have some sort of conceptual framework that is demonstrably valid for the purpose at hand. The speech, hearing and artificial intelligence sciences have always been concerned with charting the communication protocols. What I have tried to do here is to apply to its extreme an engineering-inspired way of thinking about voice, in the hope that doing so might generate new ideas in the mind of the reader. The acoustic waveform does mirror the functioning of the voice organ, neither exclusively nor completely, but in certain ways, and only if we understand not only the biological hardware but also the communication protocols in the biological software.

A similar approach could be taken for song, but the outcome would be different in some places. In the performing arts, for instance, the performer can be seen as a proxy 'sender' for the author and/or composer; but that extension was omitted here. Prosody in song is largely replaced by the music, so if the performer wants to modify points of emphasis in the message, other 'channel codes' may have to be used. In some genres of song, the 'verbal stream' is less important than in others; and so on.

Most intriguingly, recent neurophysiological research has shown that we often perceive speech and voice sounds in terms of how we would have produced the same sound with our own voice. For example, when listening to a hoarse voice, one can experience vocal discomfort without actually voicing. The sound of someone else's tense voice activates not only the corresponding motor centres in the listener's brain, but even some of the muscle motor fibres in her voice organ. Such observations have given support to 'motor' and 'mirror' theories of perception (10), which propose that in some fashion the sending apparatus is integrated or at least very well connected with the receiving apparatus. If this is so, then perhaps the task of unravelling the protocol layers of voice communication will be somewhat simplified.

Acknowledgements

The author is grateful to Rolf Carlson, Johan Sundberg, Jan Švec and David Howard for valuable discussions and feedback; and to the organisers of PEVOC 6 for the invitation to present this paper as a keynote address. Funding of relevant research projects has been provided by the Swedish Council for Working Life and Social Research (FAS), and the Swedish Foundation for Cooperation in Research and Higher Education (STINT).

Notes

1. Voice timbre usually has an expressive rather than a semantic function. A few languages employ timbral attributes such as

press or breathiness to convey meaning. This complication can be negotiated by declaring that, to the extent that timbre affects the semantics, it is an attribute of phonemes..
2. At this level, we do not separate the phonemes according to their phonetic subclasses or their acoustic features; they are all viewed as symbols in the script of the semantic stream..
3. A single value, if we are dealing with monophonic signals. Spatial sound information would require more than one acoustic channel, but this seems to be of little relevance, so long as a voice can be approximated by a point source..
4. The word 'level' is uncomfortably overloaded, in that it has precise but different meanings in acoustics, in factor experiments, in physiology, and so on. In scientific papers on voice, several of these meanings are known to have occurred in the same paragraph; the reader is herewith cautioned. In acoustics and in telecommunications, a level is the logarithm of a *ratio* of an observed power to a reference power. The level is expressed in decibels; see any textbook for details. The word 'intensity', too, has a technical meaning in acoustics (power per unit area: watts per square meter), but unfortunately this meaning is rarely upheld in the voice and speech literature. The word intensity has variously been used as synonymous with SPL, or loudness, or vocal effort, which are all different things..
5. I write the short-term average here, because while most methods do not localize the precise moments of glottal excitation, the sum of period times obtained over a running time window is usually quite accurate..

References

1. Traunmüller H. Conventional, biological, and environmental factors in speech communication: A modulation theory. Phonetica. 1994;51:170–83.
2. Quast H. Automatic Recognition of Nonverbal Speech: An Approach to Model the Perception of Para- and Extralinguistic Vocal Communication with Neural Networks. Thesis. University of Göttingen; 2001.
3. Traunmüller H. Evidence for demodulation in speech perception. Proceedings of the International Conference on Speech and Language Processing 2000. Available from: http://www.ling.su.se/staff/hartmut/demod.pdf.
4. Erman LD, Hayes-Roth F, Lesser VR, Reddy DR. The Hearsay-II Speech-Understanding System: integrating knowledge to resolve uncertainty. ACM Computing Surveys. 1980; 12(2):214–53.
5. Paliouras G, Bree DS. Adaptive Event Recognition with the use of Limited Training Data. In: Mastorakis NE, editor. Recent Advances in Information Science and Technology. World Scientific; 1998. p. 225–32.
6. Eriksson A, Traunmüller H. Perception of vocal effort and distance from the speaker on the basis of vowel utterances. Perception & Psychophysics. 2002;64(1):131–9.
7. Södersten M, Ternström S, Bohman M. Loud speech in realistic environmental noise: phonetogram data, perceptual voice quality, subjective ratings and gender differences in healthy speakers. J Voice. 2005;19(1):29–46.
8. Pinto NB, Titze IR. Unification of perturbation measures in speech signals. J Acoust Soc Am. 1990;87(3):1278–89.
9. Holmberg EB, Ihre E, Södersten M. The phonetogram as a tool in clinical voice work. Proc 6th Pan-European Voice Conf (PEVOC 6). London, 31 Aug–3 Sept 2005; ISBN 1-905351-01-1. p. 135 (abstract only).
10. Scott SK, Johnsrude IS. The neuroanatomical and functional organization of speech perception. Trends in Neurosciences. 2003;26(2):100–7.