Human Voice – a Sparse, Meaningful and Capable Representation of Sounds

R. Mores

HAW University of Applied Sciences Hamburg, Germany, Email: mores@mt.haw-hamburg.de

Abstract

This paper encourages the use of human voice features for mid-level representations of sounds. The human voice is a perfectly trained sound reference, combining two major classes of commonly used verbal sound descriptions, body tactile experience and every-day listening experience. Unlike pure technical representations, the human voice contains multidimensional physical and semantic features. It directly recalls cognitive patterns, and these, in return, bias perception. Examples of very sparse and capable representations are given by the extraction of vowel quality and nasality from short steady-state violin sounds. Such mid-level features can be identified with the help of automation, for instance, using psychoacoustic signal processing and feature extraction or using learning methods and classification. However, features can also be identified without automation, simply by listening tests and verification against a listener's own vocal tract. This general approach also supports manifold translation: backward to physical properties of a musical instrument and forward to other research fields such as cognitive musicology, or ethnological musicology, where researchers consider language-sensitive perception of sounds.

Introduction

We propose the use of human voice features for mid-level representations of short steady-state sounds. This proposal aims at bridging two stand-alone practices of describing sound. On the one hand, composers, musicians, makers of musical instruments, and recording engineers share a common language for describing sounds. Such language is more or less agreed on within communities. Even though language is not precise there is a clear image of how a flamenco guitar or a noble solo violin should sound like. On the other hand, the community of engineers prefers technical representations and automated feature extraction when analysing sound sources or when gaining metadata. Such technical representations may well be precise and repeatable, but the related numerical results and data plots do not easily translate, they cannot be listened to, nor do they create a sound image, and - even worse - they lack in representing what humans really perceive. These two practices seldom cooperate with each other in achieving descriptions that are at the same time stable and intelligible.

Language alone is not really stable and precise. Some works systematically explored the general capability of a specific language for describing sounds. Anneliese Liebe investigated 1600 German sound describing words and their use in literature from the 16th to the 19th century. She concluded that the language did not develop clear, distinct descriptions - even widely used words like 'sound' or 'tone' remain imprecise [6].

Technical representations, such as transforms, model parameters, and decompositions do only partially cover psychoacoustic reality and do not bridge to human's perception. There is limited confidence even in those commonly used examples of translation between semantic and technical descriptions: roughness, sharpness, brilliance, and loudness, to mention a few [2], [3], [4], [5]. Even today's best practice neuro-computer-science perception models yield results only at the level of pattern recognition or rough classification, and not yet at the semantic level of sound perception [1].

Searching for intelligible translations, we start at one of the major findings of Liebe: verbal sound descriptions are usually derived from other sensory experience, tangible textures, body action, visible impressions, or from comparable everyday sounds. Humans seem to search for commonly experienced reference while differentiating sounds [6]. Human's semantic sound descriptions strongly incorporate the capability of translation, i. e. the feature 'noble' learned in the strings domain can be imagined for piano sounds. A learning machine would not play such a creative role. It would stick to recognition mode and therefore require new training data before resuming work.

Voice plays an important role among the various classes of sounds such as musical instruments, every day sounds, or noise. Humans not only enjoy to imitate each other, but also to imitate other sound sources. This maybe one of the reasons why musicians refer to descriptors usually assigned to human voice, such as 'singing', 'bawling', 'chirping', or 'nasal'. Obviously, within the reference sound library used for communication, descriptors with a relation to one's own body experience seem most helpful, especially those related to the best-trained sound source: the human voice.

Dimensions of human voice

Within the feature space of human voice, physical properties are likely to be captured easily. Such low-level features are pitch, loudness, spectrum and formants, vibrato, air flow through mouth and nose, or directivity. Mid-level features such as vowel quality, nasality, articulation, or prosody are more difficult to extract. Is the vocal tract under pressure or is it relaxed? Is somebody singing or speaking, or even in a mode in between: sprechgesang? The semantic level reaches both ends, the musical and the literal dimension: rhythm, musical line, speed and acceleration, dynamics, irony, wit. Every child easily senses excitement, joy, tension, sadness or frustration in a voice, even if it hears a voice for the first time in life. Semantic features exceed the scope of this paper of finding representations for short steady-state sounds, especially since the time axis becomes more and more relevant.

Beyond this analysis approach of describing specific voice signals in the first place, human voice samples and their related feature sets incorporate the additional capability of

translation. With the use of human voice signals, it is possible to directly access a listener's resource of trained sound images. Vocal imitation of a trotting horse will recall the image of a horse and the real sound of the horse shoe. While sound descriptions on the basis of vocal imitation might be far from adequate in terms of acoustical representation, they still work, simply because a listener with similar hearing experience will recall patterns or will even construct what is missing. This translation capability goes beyond classification and works on the semantic level, too. One reason for this capability may be the early training phase, when we all learn our mother language just by listening and imitating.

The same capability is even referenced to when describing sounds verbally. What do musicians specifically mean when they say they like the singing in their cello? Can a mid-level feature "singing character" be technically extracted and correspond to what people hear?

Example: vowel quality

Here we propose the vowel quality as an intelligible sound descriptor. While playing a violin, it is a simple exercise to imitate the observed sound with the vocal tract. In many cases, one will easily identify a matching vowel, and, at the same time, a matching contribution of nasal components. Parameters may vary from semitone to semitone, however, even a child is able to identify matching vowels. The identified vowel quality is an intelligible descriptor of sound. Where musicians' verbal descriptions would vary from 'dark' to 'sonorous' for a given sound, a good portion of what they actually want to say might be expressed by the vowel /a/, in a rather precise form. Descriptions of a 'sharp' or 'shrill' component in a sound might be covered by the vowel /i/, in a compact way. Automated feature extraction could now add stability to this capable representation, and, at the same time, translate between perception and technical representation.

Searching for vowel quality in short- steady-state violin tones, Müller developed an extraction tool along selected phonetic libraries. This tool extracts the vowel quality from the formant structure in the sound and maps to the continuous space of tongue height and backness, as illustrated by the Jones diagram, Figure 1 [7]. The precision goes well beyond those rough estimates of tongue backness used in speech-to-text systems and is verified against libraries of the International Phonetic Association [8]. The development work is fully documented in [9].

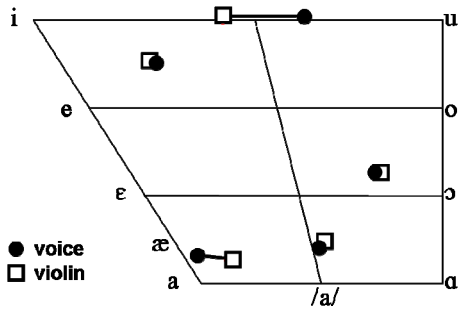

Figure 1: Examples of well matching vowel qualities in voice / violin sound pairs

The tool not only works on vocal sounds but also on strings. In a test, violin sounds have been imitated by listeners, effectively creating sound pairs of original sound and what a listener imagines while imitating the sound. The tool extracted a matching vowel quality for most of the sound pairs. Figure 1 shows some examples of sound pairs and the reader is invited to listen to the samples hosted under [10].

There are limits to this approach, which are clearly outlined in [9]. Due to the fact that the violin clearly differs from the vocal tract in terms of its generator principle: there may be no perceived vowel quality at all, or there may be a bistable perception of vowel quality in rare cases, or, in highly pitched sounds, the vowel quality becomes ambiguous.

However, from 120 violin sounds randomly chosen from various libraries 40 samples were rated as obviously revealing vowel quality, another five samples were rated as very clearly revealing vowel quality. For about 40% of the violin sounds, vowel quality captures a lot of what listeners perceive in terms of sound quality.

Again, the strength of this approach is not accurateness or perfect representation. The extracted parameters are quite useless for signal reconstruction. It is the combination of intelligibility and stability that is attractive for such analyses. The proposal takes advantage of the fact, that there is a wide and common understanding of phonetics. The idea of how 'o' sounds in the word 'home' is commonly shared among many. Additionally, the vowel quality is a very compact and precise representation, and it is very stable over time. The other advantage is the option of translation to and from technical parameters: the description may likewise be found by machine extraction or by human sound comparison. To summarize, this contribution uses the existing stable sound image of vowels as an anchor for describing the perceived quality of short steady-state sounds of strings.

When adding a statistical component to extracted vowel qualities, or when mapping the extracted vowel quality over time, a general character of an instrument may be captured. For instance, an instrument with strongly varying vowel quality across typical musical lines might be perceived as lively or vivid. On the contrary, stable vowel quality might be perceived as reserved or boring.

Figures 2 and 3 illustrate the vowel quality over time for sets of Stradivary and Guarneri violins. In all samples, the same musician is playing the same theme from Bruch's G minor concert [11]. There are many similarities between traces, as we compare within the same group of string instruments, but there are also differences between violin makers and between individual instruments.

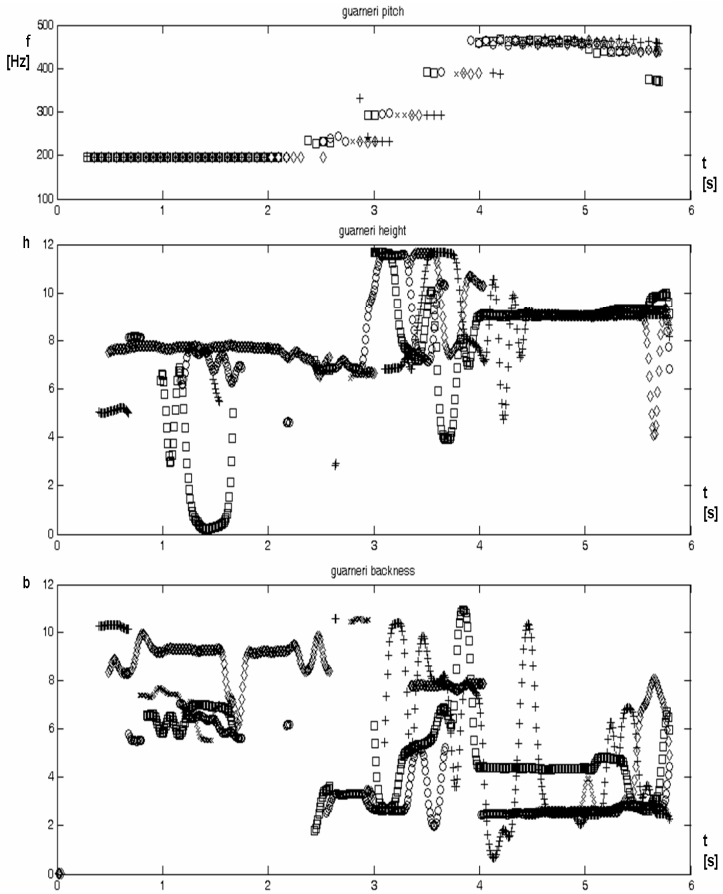

Figure 2: Vowel quality of five Guarneri violins, "Gibson" 1734 (circle), "Lafont" 1735 (cross), "Plowden" 1735 (plus), "Ex-Vieuxtemps" 1739 (square), "De Beriot" 1744 (diamond). Data cleared from out-of-range entries. Top: pitch; middle: height; below: backness.

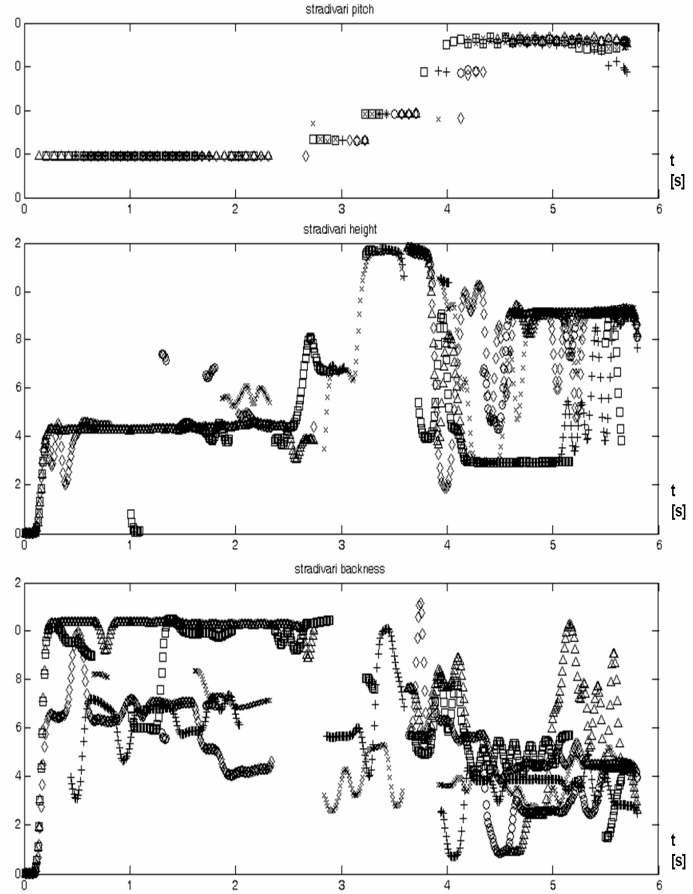

Figure 3: Vowel quality of six Stradivari violins, "Spanish" 1677 (circle), "Ernst" 1709 (cross), "Joachim" 1714 (plus), "Monasterio" 1719 (square) , "Madrileno" 1720 (diamond), "Rode" 1733 (triangle). Data cleared from out-of-range entries. Top: pitch; middle: height; below: backness.

Example: nasality

Another frequently used term is nasality. Again, such feature alone is not able to fully capture the perceived sound quality however it is one of the telling ones in a set of features. From a physical point of view, nasality has about seven ingredients observable in the frequency domain: a wider bandwidth and lower frequency of the first formant, displacement of other formant frequencies, additional resonances between 250 Hz and 500 Hz, little energy in the range of 500 Hz, additional energy between formants, less total energy of a sound [12]. Some components refer to learned sound images, others require observation over time. The difficult part in developing a "nasality meter", however, is creating an adequate perception model. Due to listeners' limited attention, components are likely to mask each other; and linear combinations seem not to be adequate. In an investigation, the typical position and the bandwidth of formants have been measured for nasal and non-nasal vocal sounds. Typical parameters for nasal sounds have been added to Stradivari sounds, and subjects have been asked for differences between original and modified violin sounds. Among the many differences perceived by individuals, nasality was not mentioned [13]. Therefore both, finding an adequate model and employing this to violin sounds seem difficult. Yet, it does not seem difficult to humans.

Sparse representation

Representations for human voice features can be very sparse. For example, the vowel quality is represented by only one byte of information for each 100 ms to 500 ms time window. The entropy is therefore 5.000 to 50.000 times smaller than the code entropy of the original 16-bit time series at sample rates between 24 kHz and 48 kHz.

Meaningful and capable representation

The capability of using human voice features as a reference is manifold. Intelligibility: the derived features can be well understood; even children are able to imitate sounds. The identified parameters directly correspond to experienced sounds. On the contrary, for most other technical features used in the engineering community there are no reference sound images that would be recalled when reading numerical results or viewing plots. Translation: representations can be extracted automatically and manually, they can directly be translated from engineering context to perception and way back. Another translation option is the back propagation of identified technical features to the physics of musical instruments. Universality: this approach bases on humans' general capability to imitate all kind of sounds with the vocal tract and also on the general readiness of recipients to understand such imitation and to construct what is missing

or to imagine the imitated feature on recalled sounds learned earlier. This is common experience, that a non-vocal sound somebody hears or imagines can directly trigger the same imagination in somebody else simply by using vocal imitation. Such carrier hosts quite a universal capability of representation. The sparse and sometimes not adequate feature representation therefore carries much more than a few bits of information. It triggers a recipient's imagination. The essence of what has been perceived and maybe communicated by means of vocal imitation will be morphed with a known target sound. In our example this target sound is the violin.

Link to ethnological musicology

Yet another capability of this simple approach is its instant usability in other specific sciences. For ethnological musicology, i.e. a valuable discussion and a research field opens up, since the rendezvous of voice and strings in the Jones diagram allows for a direct mapping of sound against the stable sound images in different languages. These images always hold discrete reference points in the diagram. Acoustically delivered vowel qualities nearby a reference point are likely to be mapped, or locked-in to that reference, as part of the cognitive process. The population of discrete reference points across the Jones plane varies from language to language. A few languages use only three distinct vowels. Languages with more than twelve vowels are relatively uncommon, although some widely-spoken languages have large vowel inventories, particularly Germanic languages. For example, the English language uses 14 to 16 vowels including diphthongs, and Swedish has the most distinct vowel qualities in the height-backness-roundedness spectrum, with 17 different monophthongs. French has 16 vowel qualities including nasals. Sedang holds the known record with 55 different vowels.

Therefore, some prejudice might go along with listeners' life-long training experienced in the given language environment. Mapping a violin's population of parameters in the Jones plane against the discrete vowel sets of specific languages might well open the question of what people really will hear.

Conclusions

The human voice is a perfectly trained sound source with the capability to imitate non-vocal sounds, too. It is therefore proposed here, to use human voice features as a reference for qualified descriptions of short steady-state sounds. Examples are the successful extraction of vowel quality from violin sounds, or nasality. A disciplined extraction of parameters that describe human voice features, will facilitate powerful representations, and the translation capability of human voice will be inherited. This approach is against the mainstream of using increasingly complex models and intense calculation, which often deliver precise but explicit results and little application value. Human voice feature representations are typically sparse and intelligible. The simplicity also facilitates application outside the engineering context and in other research fields, e.g. Ethnological Musicology.

Acknowledgements

The author thanks the Federal Ministry of Education and Research for funding.

References

[1] Adiloglu, K., Anniés, R., Wahlen, E., Obermayer, K., Representations and predictors for everyday sounds, Closing the loop of sound evaluation and design (CLOSED), Deliverable 5.1, NIPG, Berlin, 2008.

[2] Aures, W. : Der sensorische Wohlklang als Funktion psychoakustischer Empfindungsgrößen, in: Acustica, Bd. 58, S. 282-290, Hirzel Verlag, Stuttgart.

[3] Aures, W. : Ein Berechnungsverfahren der Rauhigkeit, in: Acustica, Bd. 58, S. 268-281, Hirzel Verlag, Stuttgart.

[4] Terhardt, E., Psychoakustische Grundlagen der Beurteilung musikalischer Klänge, in: Meyer, J., (Hrsg.) Qualitätsaspekte bei Musikinstrumenten, Moeck Verlag, Celle, 1988.

[5] Kloppenburg, M., Maempel, H.-J., Weinzierl, S., The appropriateness of the psychoacoustic measures sharpness and roughness for the prediction of aesthetic impression caused by sound, 24th Tonmeistertagung, Leipzig, Nov. 2006.

[6] Liebe, A.: Die Leistung der deutschen Sprache zur Wesensbestimmung des Tones, Habilitationsschrift, Berlin, 1958.

[7] Jones, D.: An Outline of English Phonetics, W. Heffer & Sons Ltd. Cambridge, 9th edition, 1962.

[8] International Phonetic Alphabet and Jones Diagram outlined in http://www.arts.gla.ac.uk/IPA/ipachart.html, access Oct. 2008.

[9] Müller, S.: Vokale in Klängen – eine LPC-basierte Extraktion der Vokalqualität zur Darstellung von Violinenklängen im Vokaldiagramm, Diplome Thesis, faculty DMI, HAW, Hamburg, Nov. 2007.

[10] Mores, R.: sound samples hosted under URL http://www.mt.haw-hamburg.de/home/mores/ , access Nov. 2008.

[11] Ricci, R.: The Glory of Cremona, Compac Disc, MCA Records, 1989.

[12] Baken, R. J., Orlikoff, R. F.: Clinical Measurement of Voice and Speech, Singular Publications, 2002.

[13] Kersten, J., Sprechen versus Singen - eine Klanganalyse an Musikinstrumenten, Diplome Thesis, faculty DMI, HAW, Hamburg, Apr. 2008.