Title

Acoustic timbre recognition

Synonyms

Sound source identification; Auditory recognition

Definition

Timbre is what allows a listener to distinguish two sounds that have otherwise the same subjective pitch, loudness, location, and duration. For instance, when orchestral musicians tune at the beginning of a concert, they all play the same note, but one can still tell the difference between instruments. This is largely because of timbre.

Detailed Description

The standard definition of timbre has several shortcomings. First, it says what timbre is not, rather than what it is. Second, it relates to the comparison between two sound tokens, whereas a more useful function for hearing is to associate a single timbre directly with a sound source (the timbre of the piano, the timbre of the voice of a friend). Perhaps as a consequence, there is still a lively debate about the acoustic features, mental representations, and neural mechanisms underlying timbre recognition. Here, we first outline the basic principles that make timbre such a powerful potential cue for sound source identification. Then we put forward two possible approaches to timbre, which we follow into the fields of acoustics, perception, neural mechanisms, and computational applications.

Why do different sound sources produce different timbres?

Sound sources are physical objects that come in all shapes and sizes. Sound is produced when some energy makes the object vibrate. The vibrations spread around the source, which then propagate to the air and reach the ear of a listener in the form of pressure waves (Figure 1). Simple physics shows that the wave pattern at the ear can contain a lot of information about what happened at the source (Helmholtz, 1877). For instance, if the energy input was brief, such as a door knock, the chances are that the sound itself will be brief and have most of its energy concentrated around the time of the knock. After the knock, the way the door continues to vibrate is closely related to its geometry, because some wave patterns are consistent with some geometries and some are not. One such rule is that waves with low frequency and thus a long wavelength are not stable within small objects. Thus, the proportions of different frequency components that combine to make the sound of a door knock will be constrained by the size of the door. Other, more complex rules apply, depending on the shape of the object, the nature of the materials involved, and so on.

Being able to decode the intricate links between wave patterns and sound sources is extremely useful for humans and other animals. It allows the auditory system to serve as a warning sense, for instance to identify sound-producing objects that are out of sight. For people, it is also the very basis of spoken language: vowels and consonants are produced by modulating the shape of the vocal apparatus, resulting in changes in timbre that are the building blocks of oral communication.

Dimensions versus features

There is no consensus on what makes timbre recognition possible for human listeners. To outline current controversies, it is useful to consider two opposite viewpoints (Figure 2). A first view is that timbre is composed of a reasonably small number of perceptual dimensions, which are subjective descriptions of sound just as pitch or loudness. Such dimensions must be metameric, in that several different sounds may project to the same point on the dimension.

A second view is that timbre recognition relies on the distinctive features of a given sound source, learnt through experience and selected amongst a very large space of potential features. The grain of a friend’s voice may be unique, which is what allows us to recognize her instantly. Such features would be conceptually different from dimensions in that a feature does not necessarily apply to all possible sound sources; in fact, it is precisely because it is unique to only a few sources (or even a single source) that it could be efficient for recognition.

It is likely that a full account of timbre will lie somewhat in between these two simplified hypotheses. However, for clarity, we continue to contrast each approach for different aspects of timbre research.

Sound representations

To investigate timbre, it is useful to represent sound visually. Classically, this has been done with tools such as the trace of the pressure waveform over time; the spectral analysis of component frequencies through e.g. Fourier analysis; or spectro-temporal transformations such as the short-term Fourier transform or wavelet analyses. More recently, computational models that aim to mimic peripheral or central auditory processing have been suggested (e.g. Patil *et al.*, 2012).

In the "dimensions" approach, summary statistics are computed on sound representations to define what are referred to as descriptors of timbre. For instance, the center of mass of all frequency components of a sound produces a single number that is correlated with the apparent "brightness" of a sound (McAdams *et al.*, 1995). In the "features" approach, the tendency is rather to maximize the richness of the representation, by including complex spectro-temporal selectivities. Such a feature-based representation need not be orderly. It can be over-complete with thousands of partially overlapping features, or sparse, in the sense that a given sound would only activate a small number of features within that large possible space (Hromadka and Zador, 2009).

Perceptual data

The basic aim of the dimensions approach is to uncover the nature and number of the perceptual dimensions underlying timbre. To this effect, statistical techniques based on multidimensional scaling have been used: a pair of sounds is presented to the listener, who has to rate how similar to each other the two sounds seem. This is repeated for all possible pairs within a given sound set. Then, the similarity judgments are treated as perceptual distances and used to obtain the dimensionality and geometry of the corresponding mental representation. For musical instruments, classic studies point towards two to three main dimensions: one related to the attack time, one related to the spectral centre of mass, and one additional dimension that is less consistently observed (Grey, 1977; McAdams *et al.*, 1995). More recent investigations, using both multidimensional scaling and verbal descriptions, suggest five main dimensions with more complex interpretations (Elliott *et al.*, 2013).

In the features approach, the focus is not on similarity but rather on the recognition of the sound source. Again using musical instruments, fast recognition times have been observed (Agus *et al.*, 2012) and recognition was found to be preserved even for severely impoverished signals (Suied *et al.*, 2013). Moreover, recognition was faster and more robust for highly familiar sources such as the human voice, an observation that could not be traced back to simple acoustic dimensions (Agus *et al.*, 2012). These results strongly suggest the existence of diagnostic features that were learnt by listeners, through experience, to recognize e.g. voices in a robust and efficient manner.

Neural bases

Neural correlates of generic timbre dimensions have been investigated with brain imaging. Using an EEG paradigm to probe sensory memory known as mismatch negativity, it has been found that timbre dimensions such as brightness or onset time could each be represented separately within auditory cortex (Caclin *et al.*, 2006).

From the features perspective, single-unit recordings have uncovered a rich variety of selectivities, at many levels of the auditory system, often without any obvious ordering principle (other than by frequency). Using linear analysis techniques such as reverse correlation, spectro-temporal receptive fields have been derived. Various spectral and temporal modulation preferences have been observed e.g. in primary auditory cortex (Depireux *et al.*, 2001). Adding a nonlinear component to the analysis adds another layer of complexity (Machens *et al.*, 2004). Furthermore, the neural encoding of timbre may interact with supposedly independent sound characteristics, such as pitch or location (Bizley *et al.*, 2009).

A further question is whether the identity of a source will be encoded by the activity of a wide network shared by many sound sources, or by the activity of only a small network specifically tuned to that source category. Evidence has been put forward for both models. Using fMRI, the identity of a sound source can be inferred from distributed activity (Staeren *et al.*, 2009). At the same time, there are clear indications of localized brain areas specialized for familiar sound sources such as the human voice (Belin, 2006).

Timbre recognition by machines

There are several applications for acoustic timbre recognition, such as speaker identification or music information retrieval. Even though the techniques used are fast-evolving and a detailed description is beyond the scope of this section, it is interesting to note that the dimensions vs. features contrast can also be seen in the architectures of the computational systems.

Automatic speech recognition, which can to some extent be viewed as a timbre-decoding exercise, has a long tradition of performing classification on a small number of generic coefficients (e.g. mel-frequency cepstrum coefficients and their variants, Hermansky, 1990). For musical instruments, a descriptors-based approach has been directly inspired by the perceptual dimensions of multidimensional studies, with a reasonably small number of explicit descriptors (Peeters *et al.*, 2011). However, other systems exist that are based on feature generation from a huge potential feature space, followed by *ad hoc* selection for a given classification task (Coath and Denham, 2005; Pachet and Roy, 2009). For musical-instrument classification, machine-learning algorithms applied on a high-dimensional auditory model representation have also been successfully demonstrated (Patil *et al.*, 2012).

Perspectives

The outstanding issues for timbre research will probably benefit from considering the various strategies available to a listener. For instance, when asked for subjective distance judgments, the most reasonable thing to do may be to abstract common dimensions to a sound set, and then use those for the comparisons. However, when asked to recognize a source as fast as possible, the mere presence of a diagnostic feature may be sufficient. The set of useful timbre dimensions or features can also depend on the task: for a same set of spoken words, different strategies are used if listeners are asked to identify the speaker or report the word content (Formisano *et al.*, 2008). Finally, the very neural representation of timbre may be dynamically tuned to the immediate acoustic context, through rapid plasticity (Fritz *et al.*, 2003). A fundamental reason that makes timbre so elusive may therefore be that timbre recognition is a profoundly adaptive mechanism, able to create and use opportunistic strategies that depend on the sounds and task at hand.

Cross-References/Related terms (optional)

Pulse Resonance Sounds; Auditory Event Related Potentials

References

Agus, T. R., Suied, C., Thorpe, S. J., and Pressnitzer, D. (**2012**). "Fast recognition of musical sounds based on timbre," J Acoust Soc Am **131**, 4124-4133.

Belin, P. (**2006**). "Voice processing in human and non-human primates," Philosophical transactions of the Royal Society of London. Series B, Biological sciences **361**, 2091-2107.

Bizley, J. K., Walker, K. M., Silverman, B. W., King, A. J., and Schnupp, J. W. (**2009**). "Interdependent encoding of pitch, timbre, and spatial location in auditory cortex," The Journal of neuroscience : the official journal of the Society for Neuroscience **29**, 2064-2075.

Caclin, A., Brattico, E., Tervaniemi, M., Naatanen, R., Morlet, D., Giard, M. H., and McAdams, S. (**2006**). "Separate neural processing of timbre dimensions in auditory sensory memory," Journal of cognitive neuroscience **18**, 1959-1972.

Coath, M., and Denham, S. L. (**2005**). "Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience," Biological cybernetics **93**, 22-30.

Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," Journal of neurophysiology **85**, 1220-1234.

Elliott, T. M., Hamilton, L. S., and Theunissen, F. E. (**2013**). "Acoustic structure of the five perceptual dimensions of timbre in orchestral instrument tones," J Acoust Soc Am **133**, 389-404.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (**2008**). ""Who" is saying "what"? Brain-based decoding of human voice and speech," Science **322**, 970-973.

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (**2003**). "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," Nature neuroscience **6**, 1216-1223.

Grey, J. M. (**1977**). "Multidimensional perceptual scaling of musical timbres," J Acoust Soc Am **61**, 1270-1277.

Helmholtz, H. (**1877**). *On the sensations of tone* (Dover, New York).

Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J Acoust Soc Am **87**, 1738-1752.

Hromadka, T., and Zador, A. M. (**2009**). "Representations in auditory cortex," Current opinion in neurobiology **19**, 430-433.

Machens, C. K., Wehr, M. S., and Zador, A. M. (**2004**). "Linearity of cortical receptive fields measured with natural sounds," The Journal of neuroscience : the official journal of the Society for Neuroscience **24**, 1089-1100.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes," Psychological research **58**, 177-192.

Pachet, F., and Roy, P. (**2009**). "Analytical features: a knowledge-based approach to audio feature generation," EURASIP Journal on Audio, Speech, and Music Processing **2009**.

Patil, K., Pressnitzer, D., Shamma, S., and Elhilali, M. (**2012**). "Music in our ears: the biological bases of musical timbre perception," PLoS computational biology **8**, e1002759.

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., and McAdams, S. (**2011**). "The Timbre Toolbox: extracting audio descriptors from musical signals," J Acoust Soc Am **130**, 2902-2916.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (**2009**). "Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex," Current Biology **19**, 498-502.

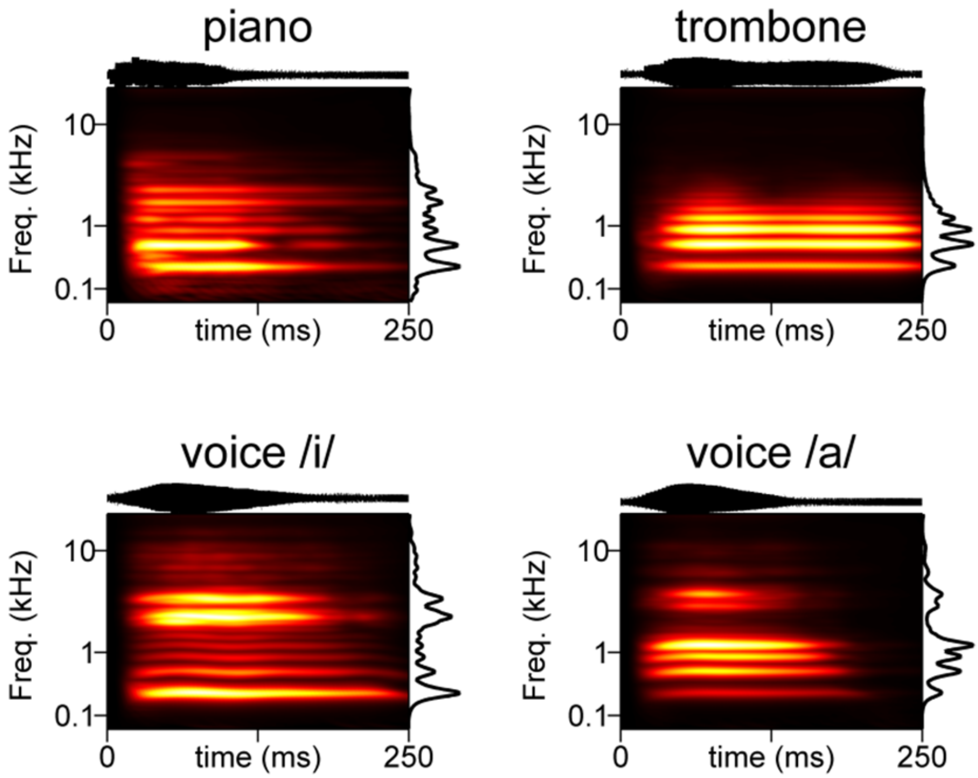
Suied, C., Agus, T. R., Thorpe, S., and Pressnitzer, D. (**2013**). "Processing of short auditory stimuli: The Rapid Audio Sequential Presentation paradigm (RASP)." in *Basic Aspects of Hearing: Physiology and Perception*, edited by B. C. J. Moore, R. D. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel (Springer, New York).

Figure legends

Figure 1. Visual representations of four sounds with the same duration, loudness and pitch, so only differing by timbre. Each panel displays a time-frequency analysis derived from an auditory model (see Agus *et al.*, 2012 for details). Briefly, color indicates the pattern of energy within frequency channels (*y*-axis) as it evolves over time (*x*-axis). The top trace is the corresponding pressure waveform. The right-hand trace is the average energy over time. The two instruments illustrate classic dimensions of timbre: depending on the sound source and how it is excited, the attack time can be fast (piano) or slow (trombone); the spectral centre of mass can be high (piano) or low (trombone). The two vowels illustrate that other, possibly more complex features may also be used to distinguish e.g. vowels from instruments, or vowels from each other.

Figure 2. Schematic representation of the dimensions approach versus the features approach for timbre. A) For the dimensions approach, all different timbres can be projected in a low-dimensional space of continuous dimensions. B) For the features approach, each timbre is defined by a set of distinctive features among a very large and unordered set of possible features.

Figure 1



265

Figure 2