CHAPTER 20.2
SPEECH AND MUSICAL SOUNDS

Daniel W. Martin, Ronald M. Aarts

SPEECH SOUNDS

Speech Level and Spectrum

Both the sound-pressure level and the spectrum of speech sounds vary continuously and rapidly during connected discourse. Although speech may be arbitrarily segmented into elements called phonemes, each with a characteristic spectrum and level, actually one phoneme blends into another.

Different talkers speak somewhat differently, and they sound different. Their speech characteristics vary from one time or mood to another. Yet in spite of all these differences and variations, statistical studies of speech have established a typical "idealized" speech spectrum. The spectrum level rises about 5 dB from 100 to 600 Hz, then falls about 6, 9, 12, and 15 dB in succeeding higher octaves.

Overall sound-pressure levels, averaged over time and measured at a distance of 1 m from a talker on or near the speech axis, lie in the range of 65 and 75 dB when the talkers are instructed to speak in a "normal" tone of voice. Along this axis the speech sound level follows the inverse-square law closely to within about 10 cm of the lips, where the level is about 90 dB. At the lips, where communication microphones are often used, the overall speech sound level typically averages over 100 dB.

The peak levels of speech sounds greatly exceed the long-time average level. Figure 20.2.1 shows the difference between short peak levels and average levels at different frequencies in the speech spectrum. The difference is greater at high frequencies, where the sibilant sounds of relatively short duration have spectrum peaks.

Speech Directional Characteristics

Speech sounds are very directional at high frequencies. Figure 20.2.2 shows clearly why speech is poorly received behind a talker, especially in nonreflective environments. Above 4000 Hz the directional loss in level is 20 dB or more, which particularly affects the sibilant sound levels so important to speech intelligibility.

Vowel Spectra

Different vowel sounds are formed from approximately the same basic laryngeal tone spectrum by shaping the vocal tract (throat, back of mouth, mouth, and lips) to have different acoustical resonance-frequency combinations. Figure 20.2.3 illustrates the spectrum filtering process. The spectral peaks are called *formants*, and their frequencies are known as formant frequencies.

The shapes of the vocal tract, simplified models, and the acoustical results for three vowel sounds are shown in Fig 20.2.4. A convenient graphical method for describing the combined formant patterns is shown in Fig 20.2.5. Traveling around this vowel loop involves progressive motion of the jaws, tongue, and lips.

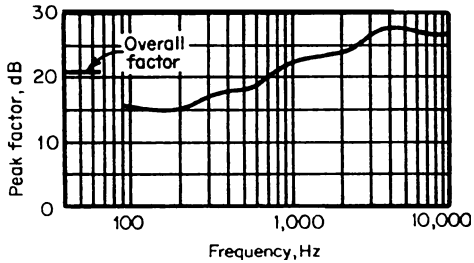

FIGURE 20.2.1 Difference in decibels between peak pressures of speech measured in short ($^1/_8$-s) intervals and rms pressure averaged over a long (75-s) interval.

Speech Intelligibility

More intelligibility is contained in the central part of the speech spectrum than near the ends. Figure 20.2.6 shows the effect on articulation (the percent of syllables correctly heard) when low- and high-pass filters of various cutoff frequencies are used. From this information a special frequency scale has been developed in which each of 20 frequency bands contributes 5 percent to a total *articulation index* of 100 percent. This distorted frequency scale is used in Fig. 20.2.7. Also shown are the spectrum curves for speech peaks and for speech minima, lying approximately 12 and 18 dB, respectively, above and below the average-speech-spectrum curve. When all the shaded area (30-dB range between the maximum and minimum curves) lies above threshold and below overload, in the absence of noise, the articulation index is 100 percent.

If a noise-spectrum curve were added to Fig 20.2.7, the figure would become an articulation-index computation chart for predicting communication capability. For example, if the ambient-noise spectrum coincided with the average-speech-spectrum curve, i.e., the signal-to-noise ratio is 1, only twelve-thirtieths of the shaded area would lie above the noise. The articulation index would be reduced accordingly to 40 percent.

Figure 20.2.8 relates monosyllabic word articulation and sentence intelligibility to articulation index. In the example above, for an articulation index of 0.40 approximately 70 percent of monosyllabic words and 96 percent of sentences would be correctly received.

However, if the signal-to-noise ratio were kept at unity and the frequency range were reduced to 1000 to 3000 Hz, half the bands would be lost. Articulation index would drop to 0.20, word articulation to 0.30, and sentence intelligibility to 70 percent. This shows the necessity for wide frequency range in a communication system when the signal-to-noise ratio is marginal. Conversely a good signal-to-noise ratio is required when the frequency range is limited.

The articulation-index method is particularly valuable in complex intercommunication-system designs involving noise disturbance at both the transmitting and receiving stations. Simpler effective methods have also been developed, such as the *rapid speech transmission index* (RASTI).

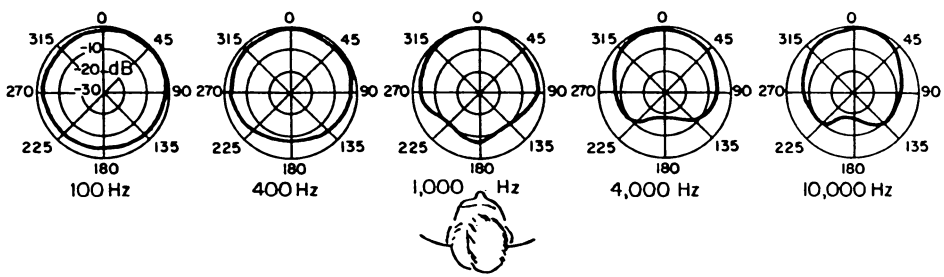

FIGURE 20.2.2 The directional characteristics of the human voice in a horizontal plane passing through the mouth.

Speech Peak Clipping

Speech waves are often affected inadvertently by electronic-circuit performance deficiencies or limitations. Figure 20.2.9 illustrates two types of amplitude distortion, center clipping and peak clipping. Center clipping, often caused by improper balancing or biasing of a push-pull amplifier circuit, can greatly interfere with speech quality and intelligibility. In a normal speech spectrum the consonant sounds are higher in frequency and lower in level than the vowel sounds. Center clipping tends to remove the important consonants.

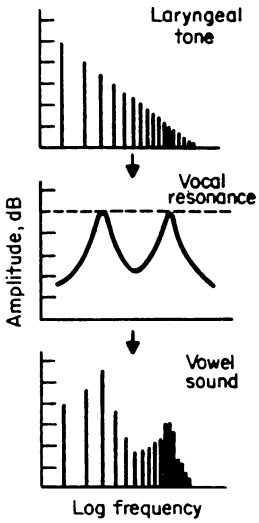

FIGURE 20.2.3 Effects on the spectrum of the laryngeal tone produced by the resonances of the vocal tract.[5]

By contrast peak clipping has little effect on speech intelligibility as long as ambient noise at the talker and system electronic noise are relatively low in level compared with the speech.

Peak clipping is frequently used intentionally in speech-communication systems to raise the average transmitted speech level above ambient noise at the listener or to increase the range of a radio transmitter of limited power. This can be done simply by overloading an amplifier stage. However, it is safer for the circuits and it produces less intermodulation distortion when back-to-back diodes are used for clipping ahead of the overload point in the amplifier or transmitter. Figure 20.2.10 shows intelligibility improvement from speech peak clipping when the talker is in quiet and listeners are in noise. Figure 20.2.11 shows that caution is necessary when the talker is in noise, unless the microphone is shielded or is a noise-canceling type.

Tilting the speech spectrum by differentiation and flattening it by equalization are effective preemphasis treatments before peak clipping. Both methods put the consonant and vowel sounds into a more balanced relationship before the intermodulation effects of clipping affect voiced consonants.

Caution must be used in combining different forms of speech-wave distortion, which individually have innocuous effects on intelligibility but can be devastating when they are combined.

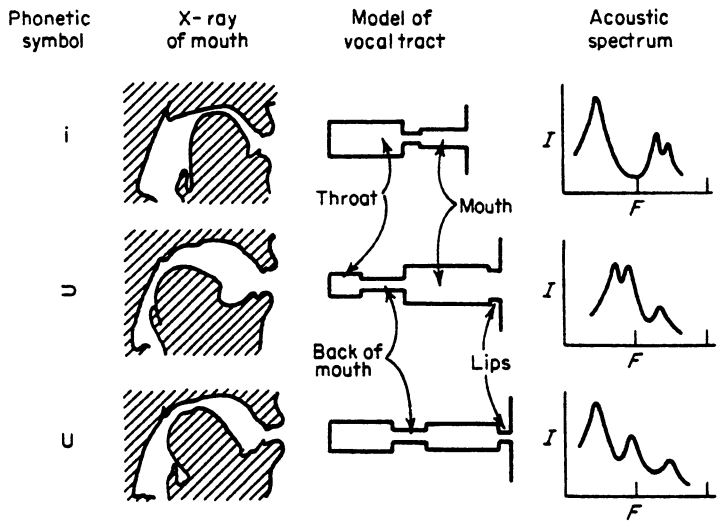

FIGURE 20.2.4 Phonetic symbols, shapes of vocal tract, models, and acoustic spectra for three vowels.[6]

MUSICAL SOUNDS

Musical Frequencies

The accuracy of both absolute and relative frequencies is usually much more important for musical sounds than for speech sounds and noise. The international frequency standard for music is defined at 440.00 Hz for A_4, the A above C_4 (middle C) on the musical keyboard. In sound recording and reproduction, the disc-rotation and tape-transport speeds must be held correct within 0.2 or 0.3 percent error (including both recording and playback mechanisms) to be fully satisfactory to musicians.

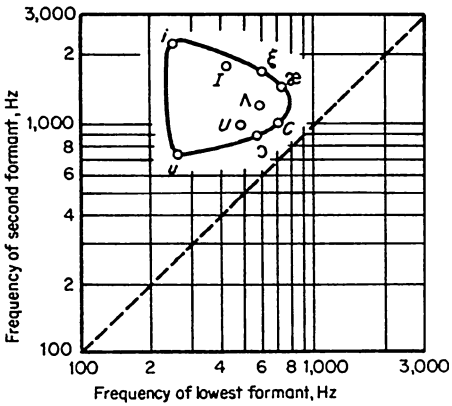

FIGURE 20.2.5 The center frequencies of the first two formants for the sustained English vowels plotted to show the characteristic differences.[7]

The mathematical musical scale is based on an exact octave ratio of 2:1. The subjective octave slightly exceeds this, and piano tuning sounds better when the scale is stretched very slightly.

The equally tempered scale of 12 equal ratios within each octave is an excellent compromise between the different historical scales based on harmonic ratios. It has become the standard of reference, even for individual musical performances, which may deviate from it for artistic or other reasons.

Different musical instruments play over different ranges of *fundamental* frequency, shown in Fig.20.2.12. However, most musical sounds have many harmonics that are audibly significant to their tone spectra. Consequently high-fidelity recording and reproduction need a much wider frequency range.

Sound Levels of Musical Instruments

The sound level from a musical instrument varies with the type of instrument, the distance from it, which note in the scale is being played, the dynamic marking in the printed music, the player's ability, and (on polyphonic instruments) the number of notes (and stops) played at the same time.

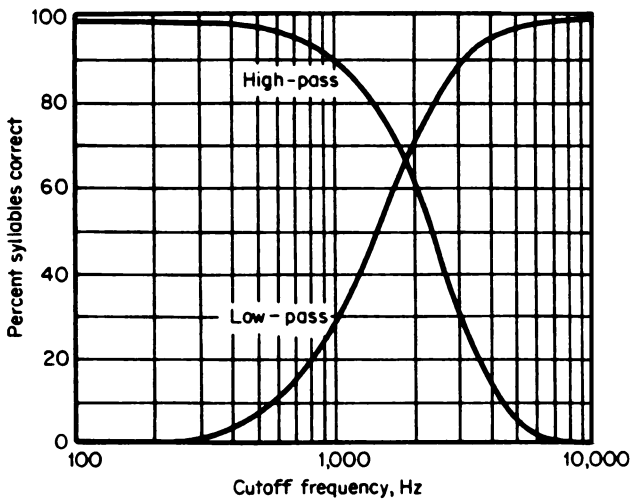

FIGURE 20.2.6 Syllable articulation score vs. low- or high-pass cutoff frequency.[8]

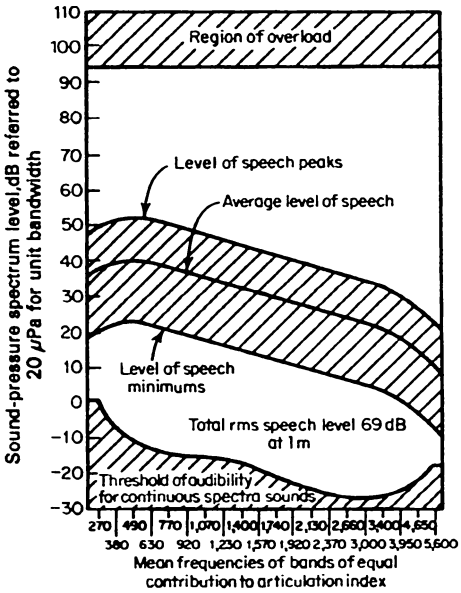

FIGURE 20.2.7 Speech area, bounded by speech peak and minimum spectrum-level curves, plotted on an articulation-index calculation chart.[9]

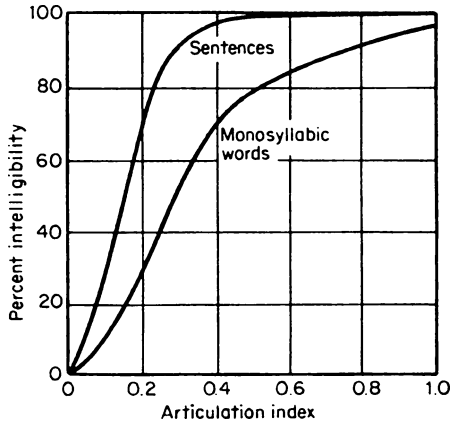

FIGURE 20.2.8 Sentence- and word-intelligibility prediction from calculated articulation index.[10]

Orchestral Instruments. The following sound levels are typical at a distance of 10 ft in a nonreverberant room. Soft (pianissimo) playing of a weaker orchestral instrument, e.g., violin, flute, bassoon, produces a typical sound level of 55 to 60 dB. Fortissimo playing on the same instrument raises the level to about 70 to 75 dB. Louder instruments, e.g., trumpet or tuba, range from 75 dB at pianissimo to about 90 dB at fortissimo.

Certain instruments have exceptional differences in sound level of low and high notes. A flute may change from 42 dB on a soft low note to 77 dB on a loud high note, a range of 35 dB. The French horn ranges from 43 dB (soft and low) to 93 dB (loud and high).

Sound levels are about 10 dB higher at 3 ft (inverse-square law) and 20 dB higher at 1 ft. The louder instruments, e.g., brass, at closer distances may overload some microphones and preamplifiers.

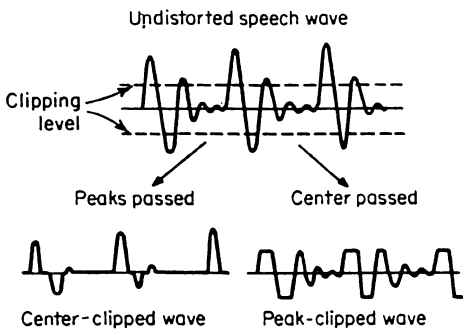

FIGURE 20.2.9 Two types of amplitude distortion of speech waveform.[5]

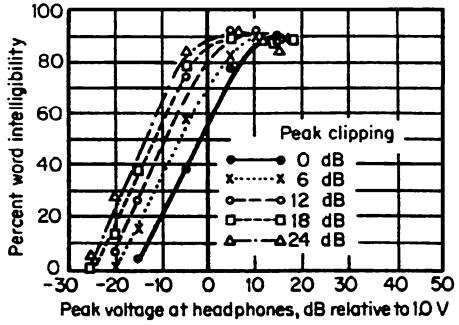

FIGURE 20.2.10 Advantages of peak clipping of noise-free speech waves, heard by listeners in ambient aircraft noise.[10]

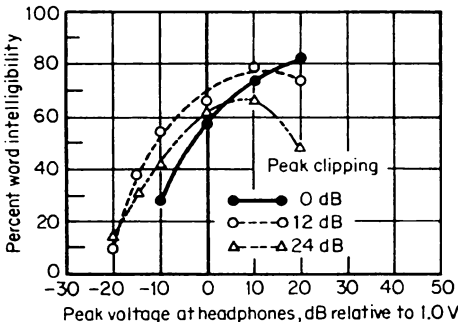

FIGURE 20.2.11 Effects of speech clipping with both the talker and the listener in simulated aircraft noise. Note that excessive peak clipping is detrimental.[10]

Percussive Instruments. The sound levels of shock-excited tones are more difficult to specify because they vary so much during decay and can be excited over a very wide range. A bass drum may average over 100 dB during a loud passage with peaks (at 10 ft) approaching 120 dB. By contrast a triangle will average only 70 dB with 80-dB peaks. A single tone of a grand piano played forte will initially exceed 90 dB near the piano rim, 80 dB at the pianist, and 70 dB at the conductor 10 to 15 ft away. Large chords and rapid arpeggios will raise the level about 10 dB.

Instrumental Groups. Orchestras, bands, and polyphonic instruments produce higher sound levels since many notes and instruments (or stops) are played together. Their sound levels are specified at larger distances than 10 ft because the sound sources occupy a large area; 20 ft from the front of a 75-piece orchestra the sound level will average about 85 to 90 dB with peaks of 105 to 110 dB. A full concert band will go higher. At a similar distance from the sound sources of an organ (pipe or electronic) the full-organ (or crescendo-pedal) condition will produce a level of 95 to 100 dB. By contrast the softest stop with expression shutters closed may be 45 dB or less.

Growth and Decay of Musical Sounds

These characteristics are quite different for different instruments. Piano or guitar tones quickly rise to an initial maximum, then gradually diminish until the strings are damped mechanically. Piano tones have a more rapid decay initially than later in the sustained tone. Orchestral instruments can start suddenly or smoothly, depending on the musician's technique, and they damp rather quickly when playing ceases. Room reverberation affects both growth and decay rates when the time constants of the room are greater than those of the instrument vibrators. This is an important factor in organ music, which is typically played in a reverberant environment.

Many types of musical tone have characteristic transients which influence timbre greatly. In the "chiff" of organ tone the transients are of different fundamental frequency. They appear and decay before steady state is reached. In percussive tones the initial transient is the cause of the tone (often a percussive noise), and the final transient is the result.

These transient effects should be considered in the design of audio electronics such as "squelch," automatic gain control, compressor, and background-noise reduction circuits.

Spectra of Musical Instrument Tones

Figure 20.2.13 displays time-averaged spectra for a 75-piece orchestra, a theater pipe organ, a piano, and a variety of orchestral instruments, including members of the brass, woodwind, and percussion families. These vary from one note to another in the scale, from one instant to another within a single tone or chord, and from one instrument or performer to another. For example, a concert organ voiced in a baroque style would have lower spectrum levels at low frequencies and higher at high frequencies than the theater organ shown.

The organ and bass drum have the most prominent low-frequency output. The cymbal and snare drum are strongest at very high frequencies. The orchestra and most of the instruments have spectra which diminish gradually with increasing frequency, especially above 1000 Hz. This is what has made it practical to preemphasize the high-frequency components, relative to those at low frequencies, in both disc and tape recording. However, instruments that differ from this spectral tendency, e.g., coloratura sopranos, piccolos, cymbals, create problems of intermodulation distortion, and overload.

Spectral peaks occurring only occasionally, for example, 1 percent of the time, are often more important to sound recording and reproduction than the peaks in the average spectra of Fig. 20.2.13. The frequency

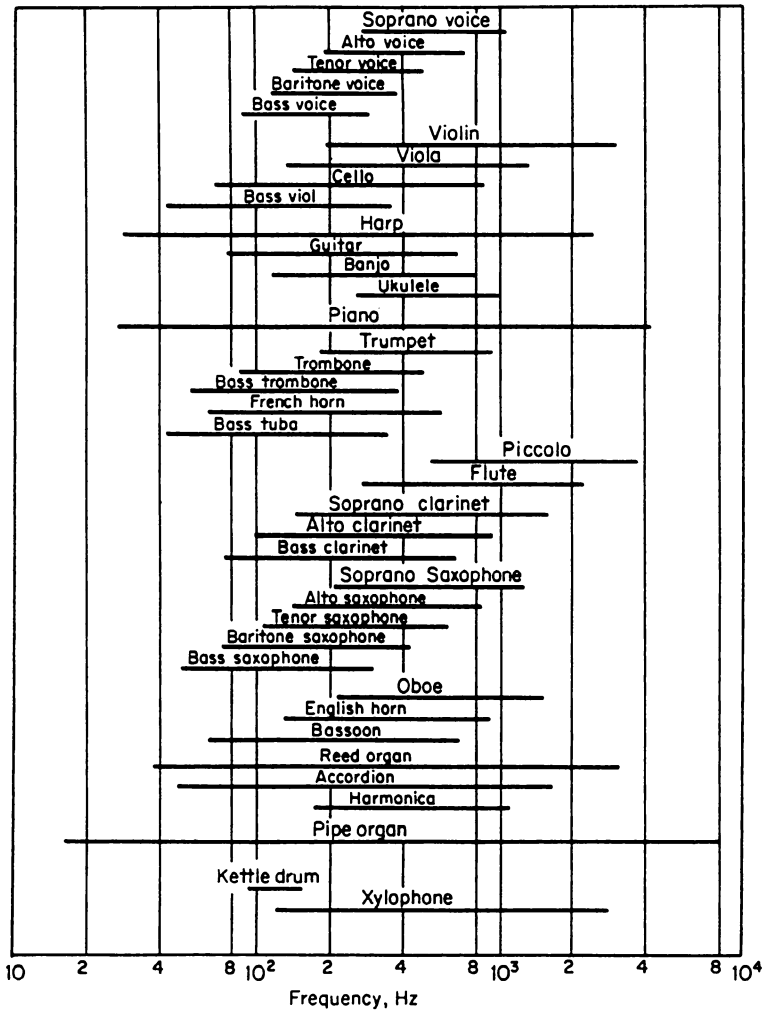

FIGURE 20.2.12 Range of the fundamental frequencies of voices and various musical instruments. (Ref. 8).

ranges shown in Table 20.2.1 have been found to have relatively large instantaneous peaks for the instruments listed.

Directional Characteristics of Musical Instruments

Most musical instruments are somewhat directional. Some are highly so, with well-defined symmetry, e.g., around the axis of a horn bell. Other instruments are less directional because the sound source is smaller than the wavelength, e.g., clarinet, flute. The mechanical vibrating system of bowed string instruments is complex, operating differently in different frequency ranges, and resulting in extremely variable directivity. This is significant for orchestral seating arrangements both in concert halls and recording studios.

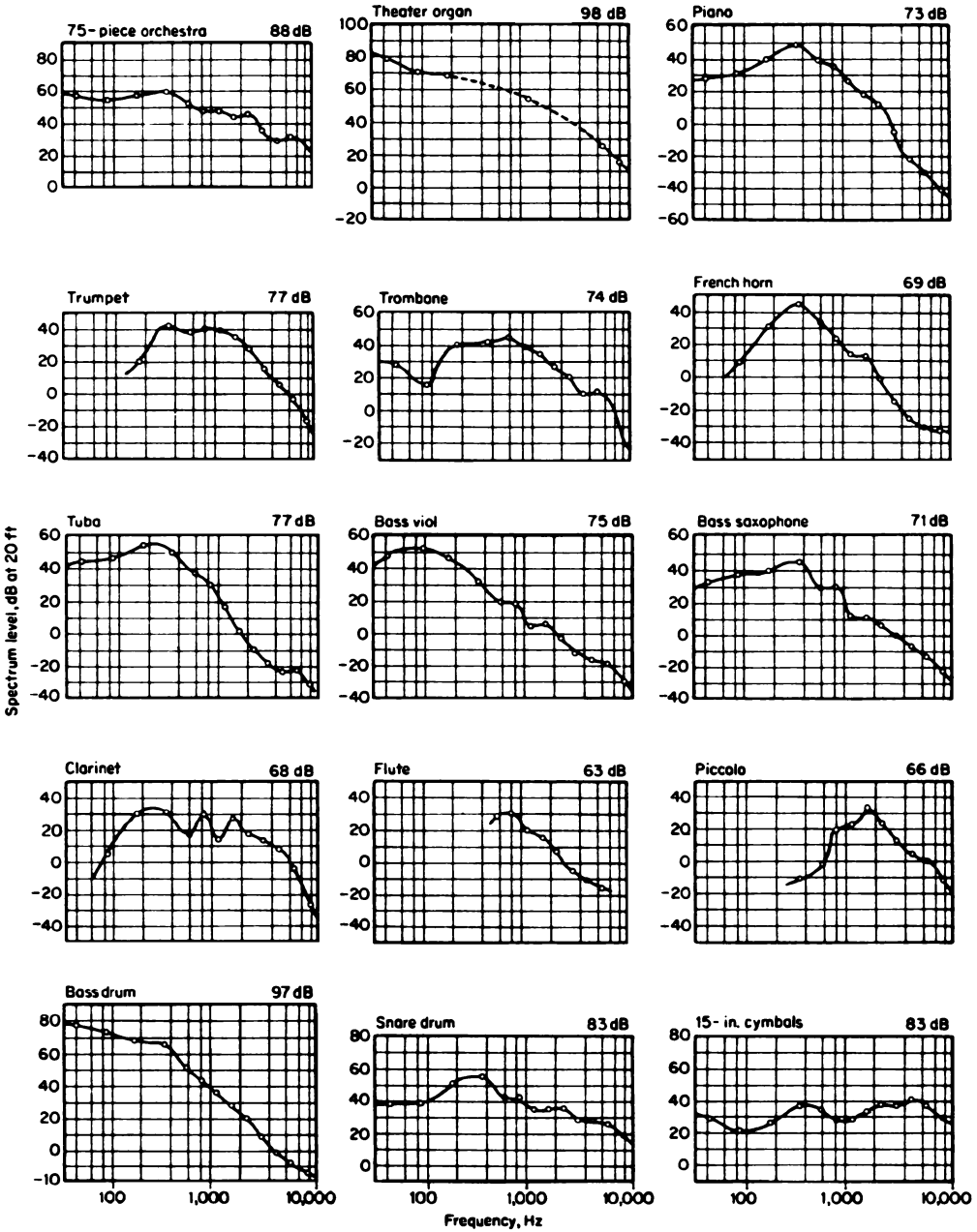

FIGURE 20.2.13 Time-averaged spectra of musical instruments.

TABLE 20.2.1 Frequency Band Containing Instantaneous Spectral Peaks

Band limits, Hz	Instruments
20–60	Theater organ
60–125	Bass drum, bass viol
125–250	Small bass drum
250–500	Snare drum, tuba, bass saxophone, French horn, clarinet, piano
500–1,000	Trumpet, flute
2,000–3,000	Trombone, piccolo
5,000–8,000	Triangle
8,000–12,000	Cymbal

Audible Distortions of Musical Sounds

The quality of musical sounds is more sensitive to distortion than the intelligibility of speech. A chief cause is that typical music contains several simultaneous tones of different fundamental frequency in contrast to typical speech sound of one voice at a time. Musical chords subjected to nonlinear amplification or transduction generate intermodulation components that appear elsewhere in the frequency spectrum.

Difference tones are more easily heard than summation tones because the summation tones are often hidden by harmonics that were already present in the undistorted spectrum and because auditory masking of a high-frequency pure tone by a lower-frequency pure tone is much greater than vice versa.

When a critical listener controls the sounds heard (an organist playing an electronic organ on a high-quality amplification system) and has unlimited opportunity and time to listen, even lower distortion (0.2 percent, for example) can be perceived.

REFERENCES

1. Miller, G. A. "Language and Communication," McGraw-Hill, 1951.
2. Dunn, H. K. *J. Acoust. Soc. Am.*, 1950, Vol. 22, p. 740.
3. Potter, R. K., and G. E. Peterson *J. Acoust. Soc. Am.*, 1948, Vol. 20, p. 528.
4. French, N. R., and J. C. Steinberg *J. Acoust. Soc. Am.*, 1947, Vol. 19, p. 90.
5. Beranek, L. L. "Acoustics,"Acoustical Society of America, 1986.
6. Hawley, M. E., and K. D. Kryter Effects of Noise on Speech, Chap. 9 in C. M. Harris (ed.), "Handbook of Noise Control," McGraw-Hill, 1957.
7. Olson, H. F. "Musical Engineering," McGraw-Hill, 1952.
8. Olson, H. F. "Elements of Acoustical Engineering," Van Nostrand, 1947.
9. Sivian, L. J., H. K. Dunn, and S. D. White *IRE Trans. Audio*, 1959, Vol. AU-7, p. 47; revision of paper in *J. Acoust. Soc. Am.*, 1931, Vol. 2, p. 33.
10. Hawley, M. E., and K. D. Kryter Effects of Noise on Speech, "Handbook of Noise Control," McGraw-Hill, 1957.