Lab 4. Acquisition and parameterization of the speech signal

Problems to prepare: speech analysis, laryngeal tone (fundamental frequency), formants, speech recognition, speaker recognition, TTS (text-to-speech)

Equipment: laptop, Reaper/Audacity, microphone, headphones

Before starting the exercise, select: sampling frequency of the recording, bit resolution, mono / stereo

Part 1:

- Recording of vowels (eg three) by each person in the group (declaring way of speaking)
- Edition of recordings: presentation of the time course of each sample separately
- Presentation of spectrograms and spectra for each sample, think about:
  - Different length of FFT (what is the smallest length of FFT to see whole formants? What is the optimum length of FFT? – too small causes small resolution however long fft requires many samples)
  - Different windowing (rectangular, Hanning, Hamming? What else?)

- Reading of the laryngeal tone frequency (fundamental frequency) (frequency in Hz and level in dBFS)
- Reading of formant frequencies (frequency in Hz and level in dBFS)
- Synth vowels using vowels-synth.pd and compare a sound of recording and synthesis
- Analysis: do you see common dependencies for a given vowel for different people and differences between individuals (frequency of the laryngeal tone, distribution of formant frequencies)

Part 2:

- Record a declarative sentence, questions, warnings
- Spectrogram analysis
- How the laryngeal tone changes over time

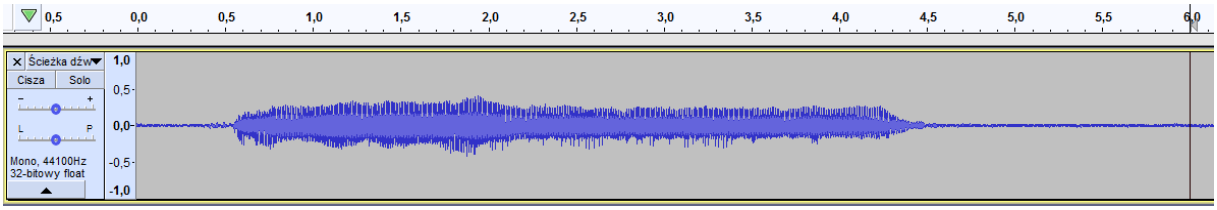
Part 3:

- using the graphic equalizer (EQ) filter different frequency bands – what can you hear? Which frequency band should be remove/gain to increase/decrease speech recognition or speaker recognition. When is it telephone sound or radio (soft lector) sound?

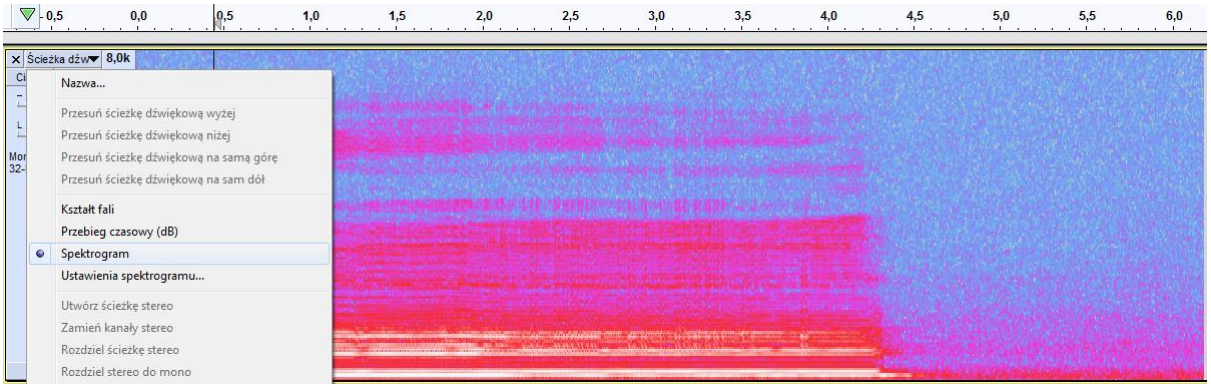
Example for part 1:

Polish vowel /a/

Time domain signal:



Spectrogram:



Spectrum:

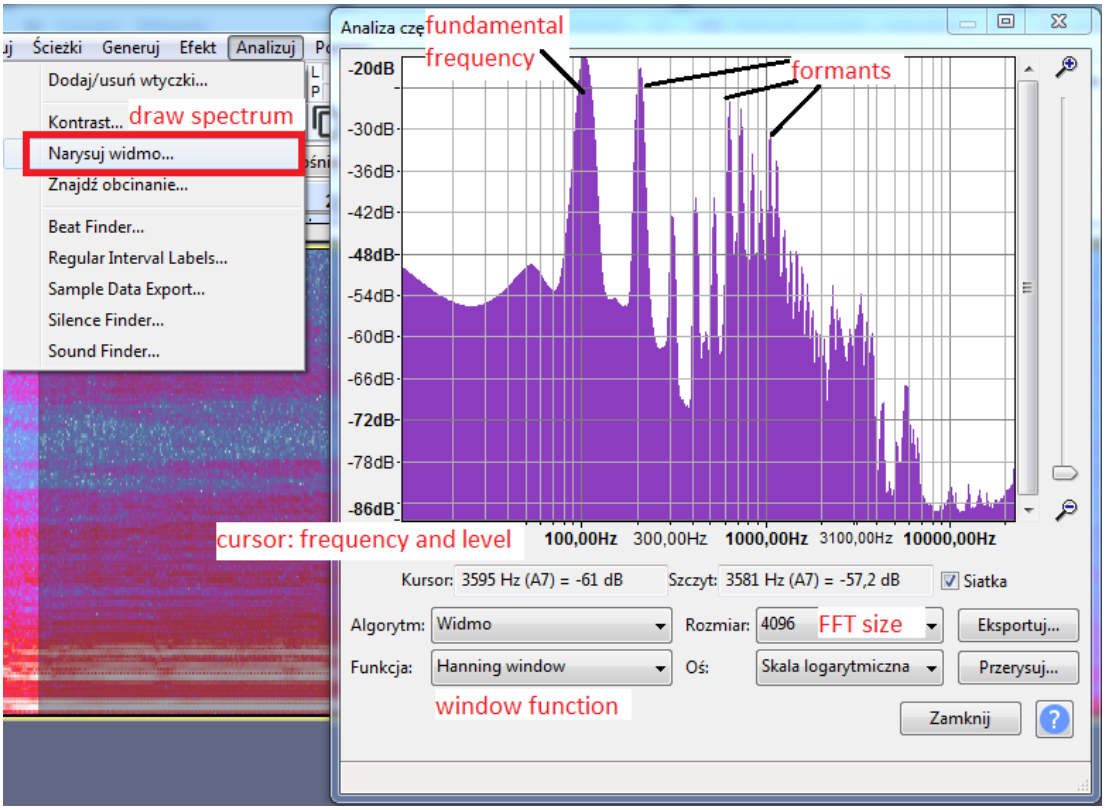
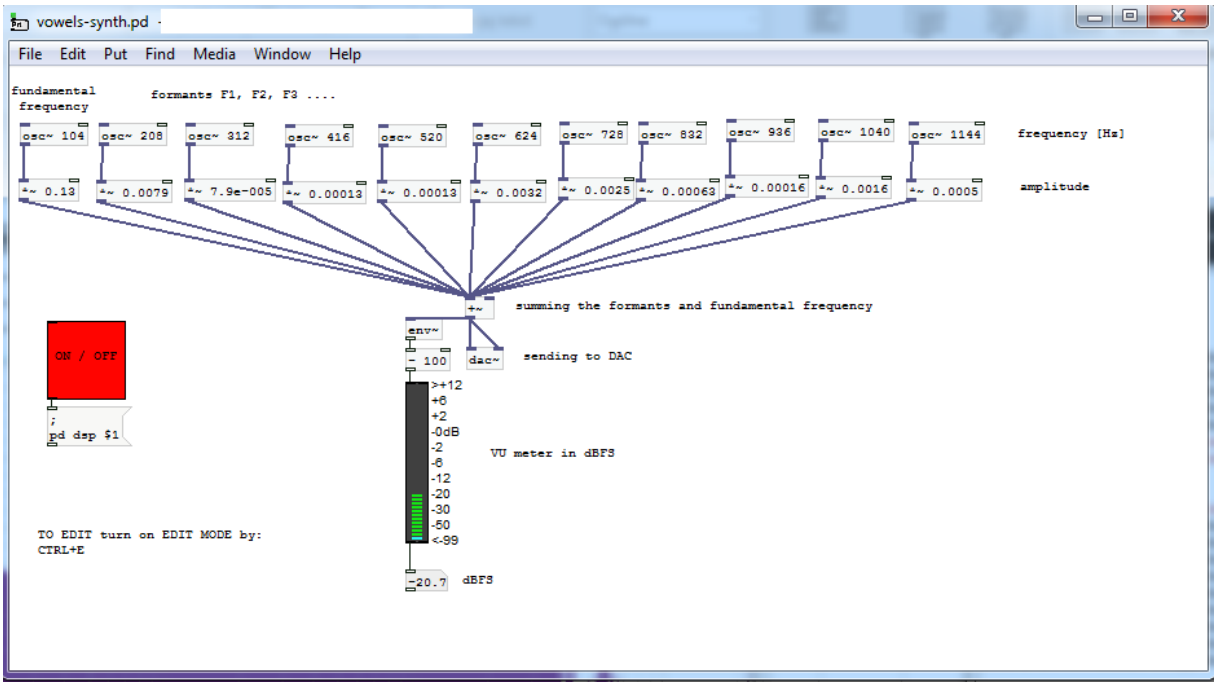


Table: Values of fundamental frequency and formants frequencies and their levels

	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	
frequency	104	208	312	416	520	624	728	832	936	1040	1144	Hz
level	-19	-21	-41	-39	-39	-25	-26	-32	-38	-28	-33	dBFS
amplitude	0,013	0,0079	0,000079	0,00013	0,00013	0,0032	0,0025	0,00063	0,00016	0,0016	0,00050	

Pure tone synthesis of polish vowel /a/



Appendix

Source: *F. Alton Everenst, Ken C. Pohlmann, Master Handbook of Acoustics, Fifth Edition, Chepter 5 Signals, Speech, Music, and Noise, pages 70 – 76*

Speech

There are two quasi-independent functions in the generation of speech sounds: the sound source and the vocal system. In general, speech is a series flow, as pictured in Fig. 5-3A, in which the raw sound is produced by a source and subsequently shaped in the vocal tract. To be more exact, three different sources of sound are shaped by the vocal tract, as shown in Fig. 5-3B. First, there is the sound we naturally think of—the sounds emitted by the vocal cords. These are formed into the voiced sounds. They are produced by air from the lungs flowing past the slit between the vocal cords (the glottis), which causes the cords to vibrate. This air stream, broken into pulses of air, produces a sound that can almost be called periodic, that is, repetitive in the sense that one cycle follows another.

The second source of speech sound is that made by forming a constriction at some point in the vocal tract with the teeth, tongue, or lips and forcing air through it under high enough pressure to produce significant turbulence. Turbulent air creates noise. This noise is shaped by the vocal tract to form the fricative sounds of speech such as the consonants f, s, v, and z. Try making these sounds, and you will see that high-velocity air is very much involved.

The third source of speech sound is produced by the complete stoppage of the breath, usually toward the front, a building up of the pressure, and then the sudden
release of the breath. Try speaking the consonants k, p, and t, and you will sense the force of such plosive sounds. They are usually followed by a burst of fricative or turbulent sound. These three types of sounds—voiced, fricative, and plosive—are the raw sources that are shaped into the words we speak.

These sound sources and signal processing can be implemented in digital hardware or software. A simple speech synthesis system is shown in Fig. 5-3C. A random-number generator produces the digital equivalent of the s-like sounds for the unvoiced components. A counter produces pulses simulating the pulses of sound of the vocal cords for the voiced components. These are shaped by time-varying digital filters simulating the ever-changing resonances of the vocal tract. Signals control each of these to form digitized speech, which is then converted to analog form.

Vocal Tract Molding of Speech

The vocal tract can be considered as an acoustically resonant system. This tract, from the lips to the vocal cords, is about 6.7 in (17 cm) long. Its cross-sectional area is determined by the placement of the lips, jaw, tongue, and velum (a sort of trapdoor that can open or close off the nasal cavity) and varies from 0 to about 3 in2 (20 cm2). The nasal cavity is about 4.7 in (12 cm) long and has a volume of about 3.7 in3 (60 cm3). These dimensions have a bearing on the resonances of the vocal tract and their effect on speech sounds.

Formation of Voiced Sounds

If the components of Fig. 5-3 are elaborated into source spectra and modulating functions, we arrive at something of great importance in audio—the spectral distribution of energy in the voice. We also gain a better understanding of the aspects of voice sounds that contribute to the intelligibility of speech in the presence of reverberation and noise. Figure 5-4 shows the steps in producing voiced sounds. First, sound is produced by the vibration of the vocal cords; these are pulses of sound with a fine spectrum that falls off at about 10 dB/octave as frequency is increased, as shown in Fig. 5-4A. The sounds of the vocal cords pass through the vocal tract, which acts as a filter varying with time. The humps of Fig. 5-4B are due to the acoustical resonances, called formants of the vocal pipe, which is open at the mouth end and essentially closed at the vocal cord end. Such an acoustical pipe 6.7-in long has resonances at odd quarter wavelengths, and these peaks occur at approximately 500; 1,500; and 2,500 Hz. The output sound, shaped by the resonances of the vocal tract, is shown in Fig. 5-4C. This applies to the voiced sounds of speech.

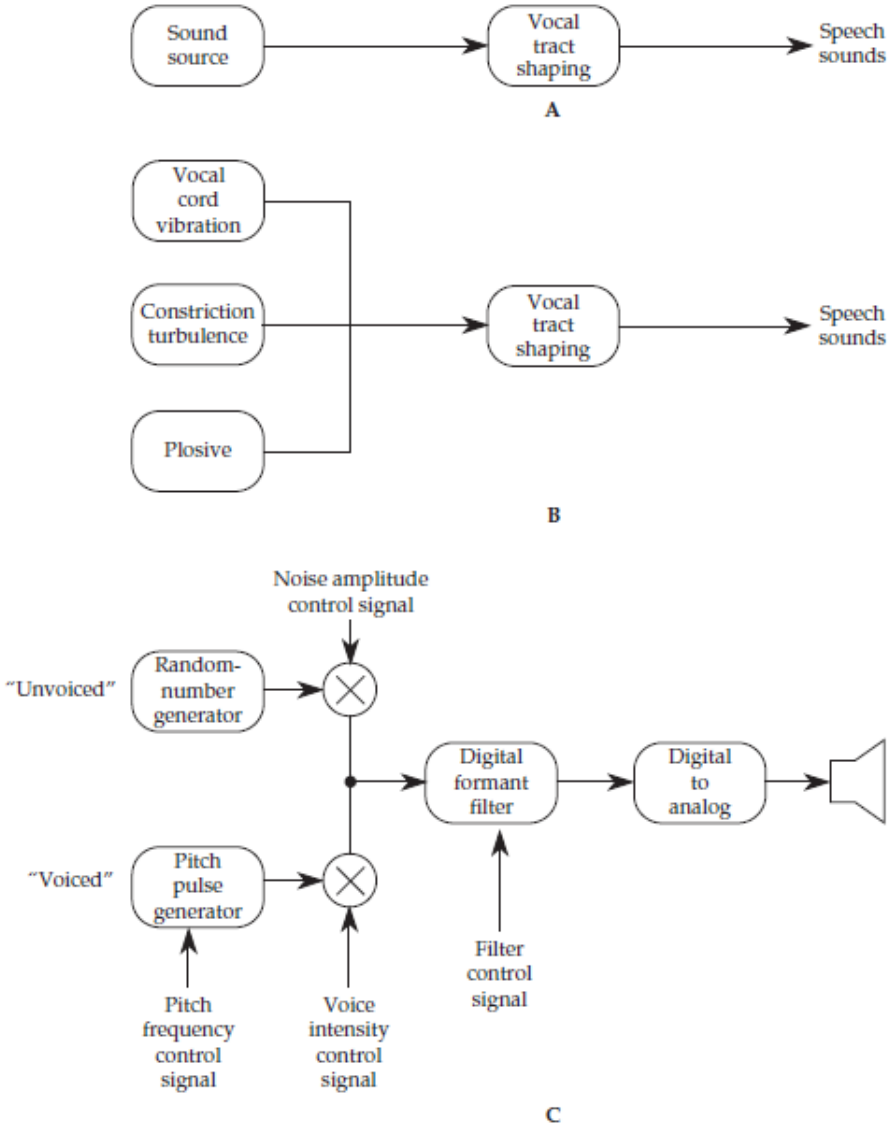


Figure 5-3 (A) The human voice is produced through the interaction of two essentially independent functions, a sound source and a time-varying-filter action of the vocal tract. (B) The sound source is comprised of vocal-cord vibration for voiced sounds, the fricative sounds resulting from air turbulence, and plosive sounds. (C) A digital system used to synthesize human speech.

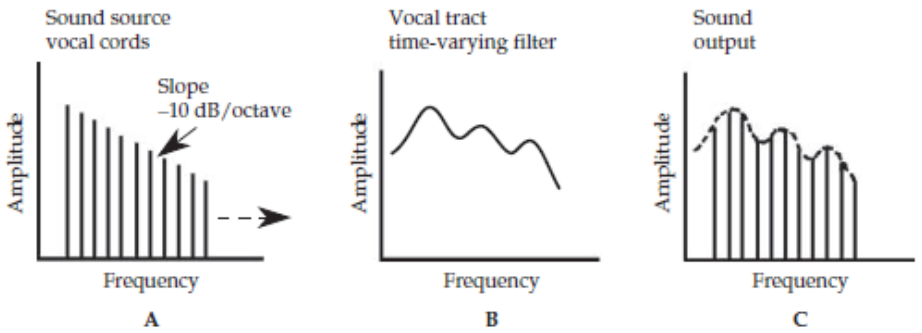


Figure 5-4 The production of voiced sounds can be considered as several steps. (A) Sound is first produced by the vibration of the vocal cords; these are pulses of sound with a spectrum that falls off at about 10 dB/octave. (B) The sounds of the vocal cords pass through the vocal tract, which acts as a filter varying with time. Acoustical resonances, called formants, are characteristic of the vocal pipe. (C) The output voiced sounds of speech are shaped by the resonances of the vocal tract.

Formation of Unvoiced Sounds

Unvoiced sounds are shaped in a similar manner, as shown in Fig. 5-5. Unvoiced sounds start with the distributed, almost random-noise like spectrum of the turbulent air as fricative sounds are produced. The distributed spectrum of Fig. 5-5A is generated near the mouth end of the vocal tract, rather than the vocal cord end; hence, the resonances of Fig. 5-5B are of a somewhat different shape. Figure 5-5C shows the sound output shaped by the time-varying filter action of Fig. 5-5B.

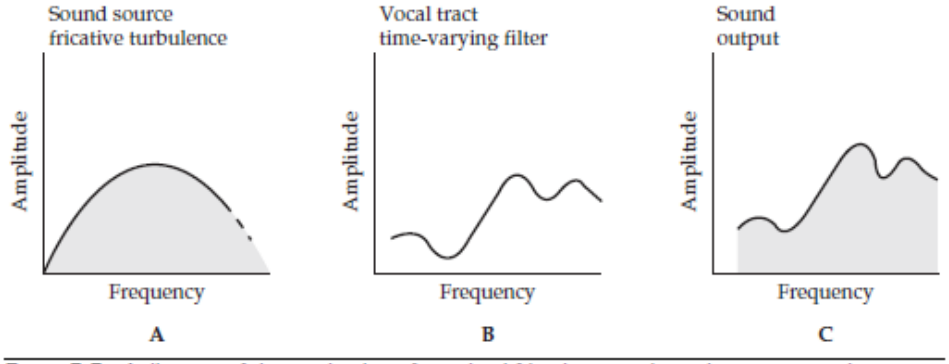


Figure 5-5 A diagram of the production of unvoiced fricative sounds such as *f*, *s*, *v*, and *z*. (A) The distributed spectrum of noise due to air turbulence resulting from constrictions in the vocal tract. (B) The time-varying filter action of the vocal tract. (C) The output sound resulting from the filter action of the distributed sound of (A).

Frequency Response of Speech

The voiced sounds originating in vocal cord vibrations, the unvoiced sounds originating in turbulences, and plosives which originate near the lips, together form our speech sounds. As we speak, the formant resonances shift in frequency as the lips, jaw, tongue, and velum change position to shape the desired words. The result is the complexity of human speech evident in the spectrograph of Fig. 5-6. Information communicated via speech is a pattern of frequency and intensity shifting rapidly with time. Notice that there is little speech energy above 4 kHz in Fig. 5-6. Although it is not shown by the spectrograph, there is relatively little speech energy below 100 Hz. It is understandable why presence filters peak in the 2- to 3-kHz region; that is where human pipes resonate.

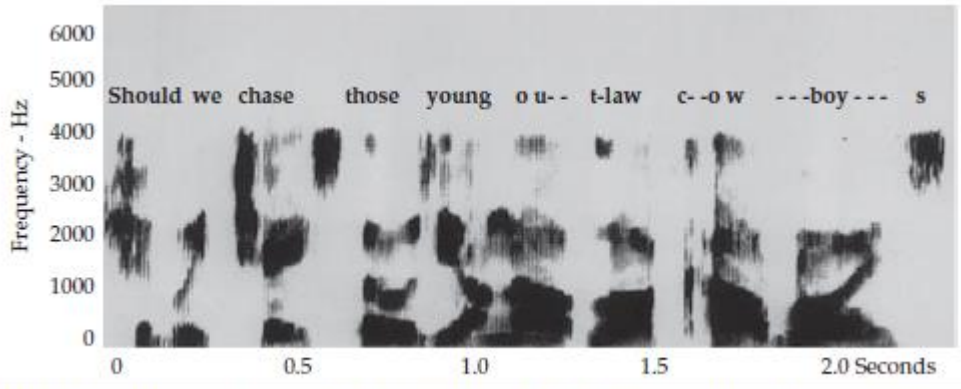


Figure 5-6 Sound spectrograph of a sentence spoken by a male voice. (*AT&T Bell Laboratories*)

Directionality of Speech

Speech sounds do not have the same strength in all directions. This is due primarily to the directionality of the mouth and the sound shadow cast by the head and torso. Two measurements of speech-sound directionality are shown in Fig. 5-7. Because speech sounds are variable and complex, averaging is necessary to give an accurate measure of directional effects.

The horizontal directional effects shown in Fig. 5-7A demonstrate only a modest directional effect of about 5 dB in the 125- to 250-Hz band. This is expected because the head is small compared to wavelengths of 4.5 to 9 ft associated with this frequency band. However, there are significant directional effects for the 1,400- to 2,000-Hz band. For this band, which contains important speech frequencies, the front-to-back difference is about 12 dB.

In the vertical plane shown in Fig. 5-7B, the 125- to 250-Hz band shows about 5 dB front-to-back difference again. For the 1,400- to 2,000-Hz band, the front-to-back difference is also about the same as the horizontal plane, except for the torso effect. The discrimination against high speech frequencies picked up on a lapel microphone is obvious (see Fig. 5-7B), although the measurements were not carried to angles closer to 270°.

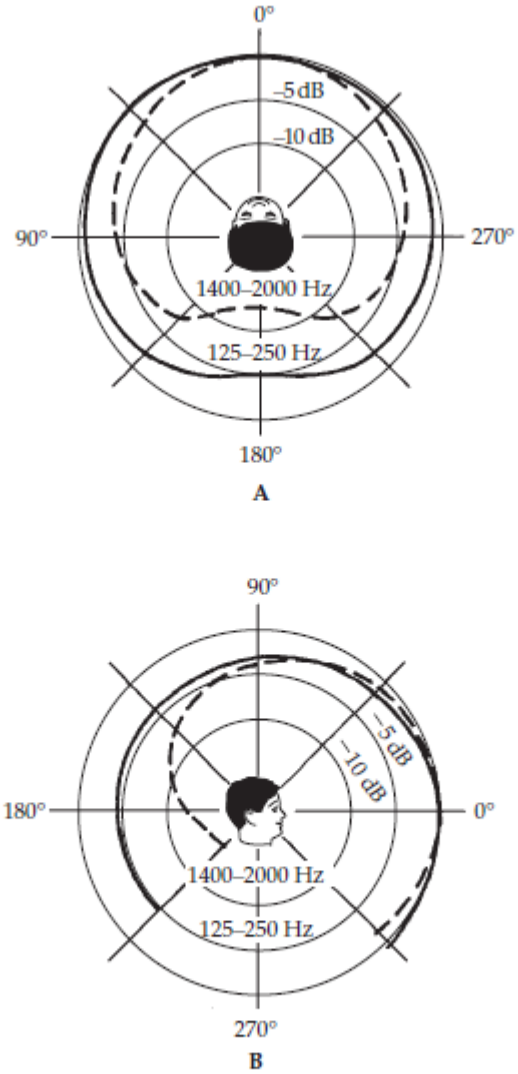


FIGURE 5-7 The human voice is highly directional. (A) Front-to-back directional effects of about 12 dB are found for critical speech frequencies. (B) In the vertical plane, the front-to-back directional effects for the 1,400- to 2,000-Hz band are about the same as for the horizontal plane. (*Heinrich Kuttruff and Applied Science Publishers Ltd., London*)