



THE EFFECT OF BANDWIDTH ON SPEECH INTELLIGIBILITY

WHITE PAPER

Jeff Rodman

Fellow/CTO

January 16, 2003

jrodman@polycom.com

© 2003 POLYCOM, INC. ALL RIGHTS RESERVED.

POLYCOM IS A REGISTERED TRADEMARK OF POLYCOM, INC. IN THE UNITED STATES AND OTHER COUNTRIES.

Abstract

Of all the elements that affect the intelligibility of speech in telephony, bandwidth has been shown to be one of the most critical. This paper discusses the relation between bandwidth and intelligibility in speech, how improved bandwidth can compensate for other deficiencies such as noise and reverberation, and how the importance of wider audio bandwidth is becoming increasingly recognized in contemporary speech communication systems.

Introduction

Some progress has been made in alleviating telephony's deficiencies in the years since the first transcontinental phone call in 1915, as many sciences have enabled a better understanding of the causes and solutions of these problems. Acoustics, physics, chemistry, and electronics enabled major advances in the design of the telephone instrument, with new designs for mouthpiece and earpiece alone producing a 10dB frequency improvement by 1940¹. Similar improvements brought closer control to the gain of these elements (early experiments required the talker to tap the carbon microphone to loosen the granules inside). As the telephone evolved, antisidetone circuits were added so the talker could better judge his own loudness. The network added echo suppression, and later, digital echo cancellation, to reduce the far-end echo that became more troublesome as long-distance calls became more routine.

However, in the last sixty years, little progress has been made in the amount of audio bandwidth that can be carried by the telephone network. Early telephone connections were not intentionally limited, but were constrained by the characteristics of the transducers and equipment then available. Intelligibility research was commonly conducted with frequencies extending from 4 kHz to 8 kHz and sometimes beyond, but the telephone network was expected to carry signals only to about 3 kHz into the 1930s, and to about 3.5 kHz with the first multiple-channel carrier systems. With standardization, and the codification of digital telephony in G.711, the upper frequency limit of the telephone network is now commonly accepted to be about 3.3 kHz at best. The last pre-divestiture Bell PSTN tests in 1984 showed significant rolloff at 3.2 kHz for short and medium connections, dropping to 2.7 kHz in long distance connections². At the low end of the spectrum, the telephone network carries frequencies no lower than 220Hz, and most commonly only as far down as 280 or 300 Hz.

In contrast to this telephone performance, we find FM radio and television spanning 30Hz to 15 kHz, CD audio covering 20 Hz to 20 kHz, professional and audiophile audio 20 Hz to above 22 kHz, and AM radio extending up to 5 kHz. Polycom's new VTX technology extends business telephone audio to 7 kHz on standard analog phone lines, and even the desktop telephone in IP systems is being considered for operation to 7 kHz.

Bandwidth and Intelligibility

Crandall noted in 1917, "It is possible to identify most words in a given context without taking note of the vowels...the consonants are the determining factors in...articulation."³ "Take him to the map" has a very different meaning from "take him to the mat," and a handyman may waste a lot of time fixing a "faucet" when the faulty component was actually the "soffet." Pole, bole, coal, dole, foal, goal, told, hole, molt, mold, noel, bold, yo, roll, colt, sole, dolt, sold, toll, bolt, vole, gold, shoal, and troll all share the same vowel sound, only differing in the consonants with which it is coupled, but the difference can drastically change the meaning of a sentence. Consonant sounds have this critical role in most languages, including French, German, Italian, Polish, Russian, and Japanese⁴. And, of course, consonants occur frequently in speech. "P" and "t," one of the most commonly confused pairs, account for over 10 percent of the phonemes in simple speech. "F" and "s" are 6.8 percent, "m" and "n" another 10.3 percent, and so on⁵. Overall, more than half of all phonemes are consonants.

This critical role of consonants in speech presents a serious challenge for the telephone network. The reason for this is that the energy in consonant sounds is carried predominantly in the higher frequencies, often beyond the telephone's bandwidth entirely. While most of the average energy in English speech is in the vowels, which lie below 3 kHz, the most critical elements of speech, the consonants, lie above. The difference between "f" and "s," for example, is found entirely in the frequencies above 3 kHz; indeed, above the 3.3 kHz telephone bandwidth entirely. Note (Figure 1) how the burst of high-frequency sound that distinguishes the "s" in "sailing" from the "f" in "failing" occurs between 4 kHz and 14 kHz. When these frequencies are removed, no cue remains as to what has been said.

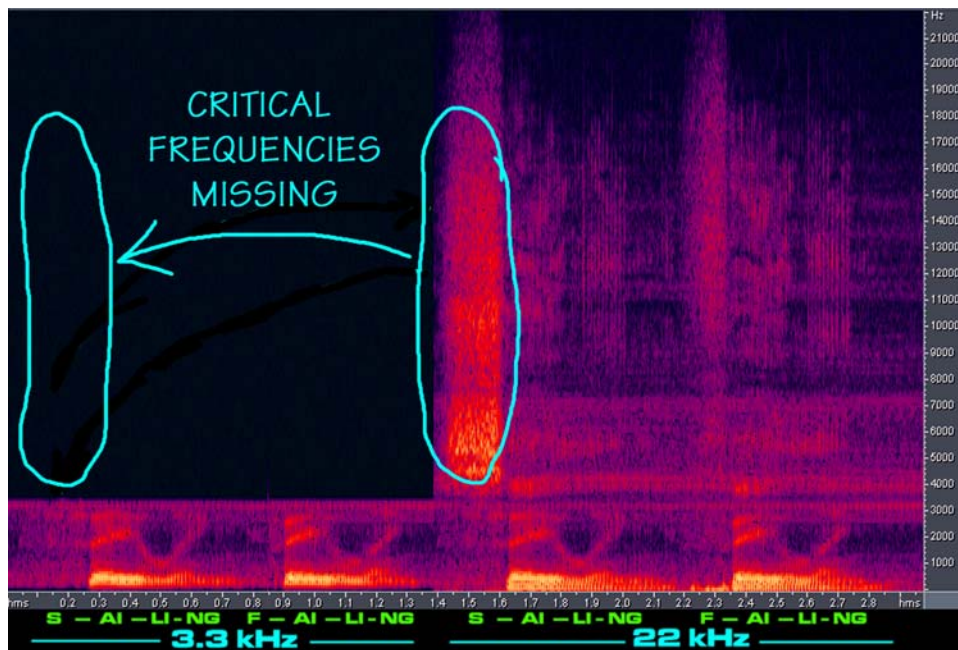


Fig 1: Speech spectra of "sailing" and "failing" at 3.3 kHz and 22kHz

This makes a conventional telephone incapable of conveying the difference between "my cousin is sailing in college" and "my cousin is failing in college" without the analysis of additional contextual information (knowing whether my cousin sails frequently, for example).

Bandwidth, Hearing, and the Mechanics of Voice

Overall, two-thirds of the frequencies in which the human ear is most sensitive and 80 percent of the frequencies in which speech occurs are beyond the capabilities of the public telephone network. The human ear is most sensitive at 3.3 kHz, just where the telephone network cuts off. The PAMS (perceptual analysis/measurement system, a standard method for measuring perceived quality of speech) rating of 7 kHz speech can be a full point (5 vs. 4) over that of 3.3 kHz speech. ⁶

The human voice uses different mechanisms to form consonants as compared to vowels. The vibration of the vocal cords, when unfiltered, is a raw buzzing sound. Vowels are produced by shaping the frequencies in this buzzing through the vocal tract, in the same manner that the plumbing in a trumpet mellows and tunes the "bronx cheer" which the player blows into the mouthpiece. The speech tract concentrates these frequencies in *formants*, which are loose groupings of frequencies occurring at roughly 400 Hz, 1200 Hz, and 2000 Hz (although formant frequencies can vary markedly, going to 3900 Hz and above ⁷).

Consonants, by comparison, are nonvoiced clicks, puffs, breaths, etc. They are created, not from the vocal cords, but by colliding, snapping, and hissing through combinations of tongue, cheeks, teeth, and so on. While formants are useful in the analysis of vowel sounds and long, voiced sounds, we see that they have very little to do with those elements of speech that carry so much of its information, the consonants.

Intelligibility, Bandwidth, and Fatigue

The most common approaches to measuring speech accuracy involve an announcer who reads lists of syllables, words, or sentences to a group of listeners. The *articulation score* is the percentage of these that are correctly recorded by the listeners. Articulation index, word articulation, and syllable articulation are all measurements that measure the degree of accuracy with which a listener can determine a spoken word.

Measurements consistently show that the intelligibility of speech decreases with decreasing bandwidth. For single syllables, 3.3 kHz bandwidth yields an accuracy of only 75 percent, as opposed to over 95 percent with 7 kHz bandwidth⁸.

This loss of intelligibility is compounded when sounds are combined in sentences. A sentence composed of ten words, each with 90 percent reliability, has only a 35 percent (0.9^{10}) probability of being understood clearly (Figure 2). In normal speech, words come

at a rate of about 120 words per minute. Consequently, 3.3 kHz speech produces about 40 ambiguities per minute, where 7 kHz speech will produce fewer than four, or close to the accuracy of live open-air speech.

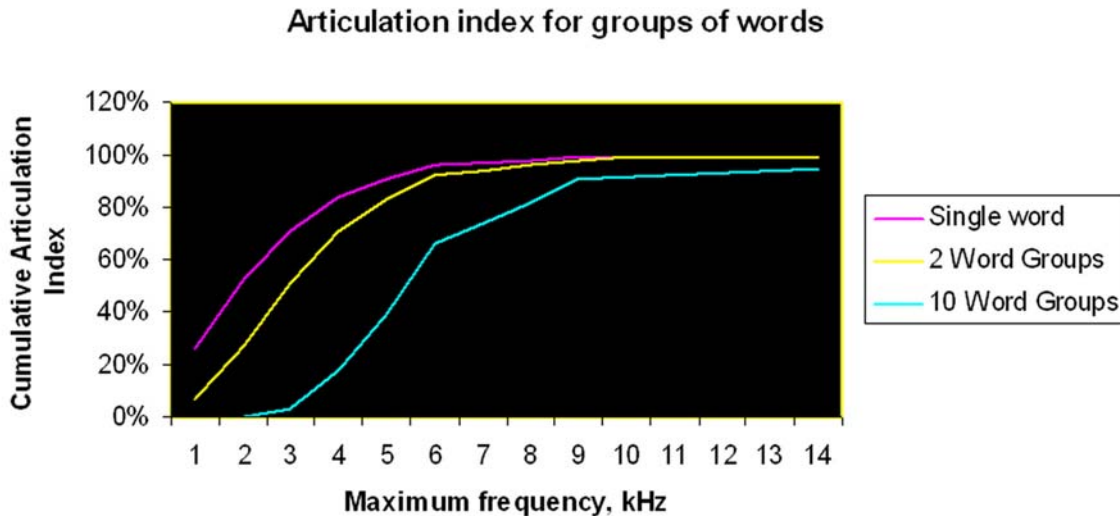


Figure 2: Low bandwidth causes lower accuracy in sentences

We are not conscious of confusion this frequently because the brain has some ability to compensate. When a sound is not clear, the brain attempts to examine the context of the sound.

The first analysis is grammatical: of the possibilities, which ones fit the grammar of the sentence? "I have to tie my choose" makes no sense, for example, so the word is probably "shoes."

When multiple possibilities fit grammatically, the listener then tries to decide what would make sense in the context of the meeting. Marine biologists may be discussing a *dolphin*, for example, while French historians are more likely to be pondering a *dauphin*.

However, when presented with a continual string of such verbal puzzles as the meeting progresses, the listener is distracted. Pieces of the conversation are lost on these mental detours, trying to deduce what words were used. As this occurs over and over in a meeting, fatigue increases, while comprehension and interaction drop. The listener has to divert her attention much more often to figuring out what words were spoken, instead of staying with the flow of the conversation. Too much of the listener's time is spent in unraveling the intended meaning, instead of understanding it.

Additional Factors Affecting Speech Accuracy

There are additional aspects of business conferencing that interact with audio bandwidth.

Reverberation

Reverberation, which comes from the natural reflections occurring in any room, magnifies the degrading effect of limited bandwidth. This is an important issue in business telephony, because group teleconferences are usually held in meeting rooms, which are reverberant spaces. This problem is also magnified as the talker moves farther from the microphone, or when the microphone is pointed away from the talker, because a larger proportion of the total received sound is reverberant rather than direct.

Compounding the problem, reverberance of rooms is magnified by audio conferencing systems because only one "ear" (the microphone) is available to the listener over the telephone link. With two ears, the brain automatically eliminates much of the reverberation (through mechanisms that are not well understood even today). But in a teleconference, only one channel is carried, so the brain is unable to provide these dereverberation functions. The next time you are listening to a talker in a large hall, try closing your eyes, and then block one ear. Notice how much harder the talker is to understand? This is why a talker may sound clear when you are sitting ten feet away, but muddy and reverberant even through a high-quality microphone placed in the same location.

Increasing bandwidth is very effective at counteracting this problem. In one test, word accuracy in a reverberant space was only 52 percent when the available bandwidth was 4 kHz; increasing bandwidth to 7 kHz raised this to 80 percent⁹.

Accented Speech

The expansion of global business has increased the importance of accurate telephone communication among talkers who have different native languages or dialects. Understanding accented speech can be much more difficult than native speech, both because of the presence of an accent, and because grammar, pronunciation, and even word selection are much different than the listener expects. A Korean speaker of English, for example, will commonly substitute "p" for "f" ("faint" becomes "paint," "coffee" becomes "copy"). A Turkish speaker may insert extra syllables ("stone" becomes "istone" or "sitone"). Even a speaker in London, referring to a cigarette container as a "fag packet" (notice all the consonants?), may leave his American listener completely perplexed.

This increases the importance of the physical parameters of speech communication (bandwidth, reverberation, amplitude, interaction, and noise) because it is no longer safe to assume that an unclear word can be deduced from its grammatical context. The backup strategies previously described assume that a shared, correct grammar was intended; when the speaker and listener have different backgrounds, this is no longer a reliable assumption. Hence, the increased accuracy that derives from increasing speech bandwidth is more critical when speech is accented.

Soft and Whispered Speech

Another variable that affects the relative importance of the higher frequencies is the use of whispered speech. While the long-term average energy at 7 kHz in normal speech is roughly 40 dB below that at 600 Hz, in whispered speech it is almost flat, dropping only 10 dB over these three octaves. Hence, in whispers, even the vowels are much less intelligible with telephone bandwidth. A person with a cold, or growing hoarse, will have more difficulty being understood both because they have proportionately less energy within the telephone band, and because they are probably speaking more softly.

Subtelephonic Frequencies

One part of the speech spectrum that bears mention is in the lower frequencies, those below the 300 Hz telephone limit. The fundamental frequency of the vowels in human speech is around 100 Hz, the frequency at which the vocal cords actually vibrate. The spectrum of the vowel "o" is shown in Figure 3. It is apparent that only a fraction of the available speech energy is present (yellow) within the telephone bandwidth.

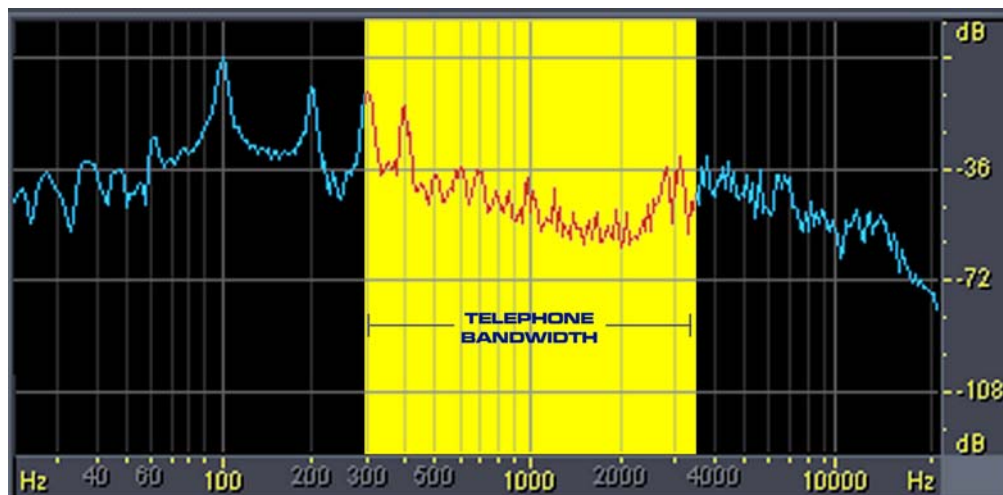


Fig 3: Vowel "o," male speaker

This is interesting because it shows that the telephone removes important frequencies both above and below its passband. Not only the higher frequencies, but also the lower frequencies are important. In general, the telephone's elimination of frequencies below 250 Hz is responsible for much of the "unreality" and loss of comfort that we hear in telephonic speech, the sense that the talker is not really present. Tests have also shown that there are significant cues in the infrasonic range (20 to 80 Hz) with consonants such as "p," "b," "k," "t," and "d"¹⁰.

Contemporary Speech Communication Systems

The problems resulting from limited bandwidth in speech communication are widely recognized today. As a result, speech communication systems with expanded bandwidth are in expanding use. In video conferencing, for example, audio connections commonly have a bandwidth of 7 kHz, and sometimes 14 kHz or higher. FM and television carry sound with 15 kHz bandwidth. IP telephony is moving to 7 kHz bandwidth using compressed and uncompressed codec techniques as described in TIA 920-200, and even the cellular telephone network is expanding to enable 7 kHz audio via the G.722.2 speech codec.

Conclusion

By extending telephone bandwidth to 7 kHz and beyond, it is clear that one can markedly reduce fatigue, improve concentration, and increase intelligibility. It is also clear that this improvement is even more significant in real-world room situations, where the sound is often degraded by reverberation, projector or air conditioner noise, accented speech, and other acoustic problems that are encountered in business telephony. Additionally, extending telephone bandwidth below 300 Hz brings a significant increase in presence and realism.

In his 1938 paper discussing the bandwidth of the telephone system, AT&T's Inglis noted that, "Frequency limitation is essentially an economic one, subject to change as conditions change."¹ Here in the twenty-first century, economics and conditions have changed as Inglis predicted, and modern telephony is now in a position to deliver on the promises of wider bandwidth and clearer speech.

¹ A. H. Inglis, "Transmission Features of the New Telephone Sets," Bell System Technical Journal 17 (1938): 358-380.

² M. B. Carey, H. T. Chen, A. Descloux, J. F. Ingle, and K. I. Park, "1982/83 End Office Connection Study: Analog Voice and Voiceband Data Transmission Performance Characterization of the Public Switched Network," AT&T Bell Laboratories Technical Journal, 63 No. 9 (November 1984).

³ I. B. Crandall, "The Composition of Speech," Phys. Rev. 10 ser. 2 (1917): 75.

⁴ John Collard, "A Theoretical Study of the Articulation and Intelligibility of a Telephone Circuit," Electrical Communication 7 (1929): 174.

⁵ P. B. Denes, "On the Statistics of Spoken English," The Journal Of the Acoustical Society of America 35 (6) (1963): 892-904 .

⁶ Anthony Rix and Mike Hollier, "Perceptual speech quality assessment from narrowband telephony to wideband audio", AES 107th Convention, New York: 24-27 September 1999.

⁷ Gordon E. Peterson, "The Information-Bearing Elements of Speech," The Acoustical Society of America Journal, 24 (6) (1952): 632 .

⁸ N. R. French and J. C. Steinberg, "Factors Governing the Intelligibility of Speech Sounds," The Acoustical Society of America Journal, 19 (1) (1947): 90.

⁹ P. W. Barnett, "Overview of Speech Intelligibility," Proceedings of the Institute of Acoustics, 21 Part 5 (1999).

¹⁰ L. L. Myasnikov, E. M. Miasnikova, and M. Y. Pikel'nyi, "Infrasonic Cues for the Automatic Recognition of Speech Sounds," Soviet Physics - Acoustics, 14 No. 4 (April-June, 1969): 522.

¹¹ A. H. Inglis, "Transmission Features of the New Telephone Sets," Bell System Technical Journal 17 (1938): 358-380.