# Telecom Technology and Audio Quality

Richard A. THOMPSON
Telecommunications Program, School of Information Sciences, University of Pittsburgh,
Pittsburgh, PA 15260, USA rthompso@pitt.edu

## ABSTRACT

While successive technology has elevated most aspects of audio quality, this paper shows how it has lowered some aspects, especially bandwidth. We show how technology and network integration reduce our ability to discriminate phonemes and identify speakers. High-fidelity Voice-over-IP is proposed as way to fix this.

**Keywords:** bandwidth, fidelity, integration, speech, VoIP

## 1. INTRODUCTION

Telecom technology has been a remarkable benefit to the human species. Morse, Bell, Marconi, and Zworykin gave us the ability to communicate over distance, but their inventions were characterized by greatly reduced aural and visual quality. During the last century, successive technology raised many aspects of the original application quality, but some technology has actually lowered other aspects. Two examples are: successive layers of technology have successively reduced audio bandwidth and the recent transition to digital television broadcast introduced an annoying pixel-block "dance" when video compressors reset after receiving noisy packets. This paper discusses the devolution of audio quality and discusses how we don't have to live with it. A later companion paper will discuss video quality.

Section 2 describes the human capacity for aural quality. Section 3 reviews the history of evolving and devolving quality, the sources and reasons that technology degrades audio quality, and the history of the complaint about this devolution. Section 4 describes the effect of *network integration* on app quality. Finally, Section 5 defines and proposes *high-fidelity Voice-over-IP*.

## 2. HUMAN CAPACITY FOR AURAL QUALITY

Since most readers probably aren't familiar with the details, this section reviews the anatomy, physics, physiology, and *brainware* of human speech and hearing, and describes how we discriminate phonemes and recognize speakers.

### Review of Human Speech

Speech is a complex acoustic signal we humans emit and receive. It is a sequence of air compressions and rarifications, which travels about 770 mph. Speaking requires a complex structure (see Figure 1A) in which, by modulating an exhaled air stream, we emit sequences of elementary sounds, called *phonemes.*

If we partly close our larynx as we exhale, our "vocal cords" vibrate at a fundamental pitch, $f_1$ = 80 to 350 Hz depending on the speaker's shape, gender, and age. By altering tension, we can change $f_1$ to any value between half and double its regular pitch – for singing and linguistic cues. Since the acoustic waveform at the larynx resembles a sawtooth, it is rich in harmonics.

Our mouth is a variable resonant cavity that acts as a tunable acoustic filter. By changing its internal shape, we alter the acoustic signal's harmonics as they pass through. Our two main techniques are to change the *tongue position* and to switch our *nasal cavity* in/out with our *uvula.* See Figure 1B. Each phoneme has a different recipe of the weights of the harmonics. See Figure 1C.

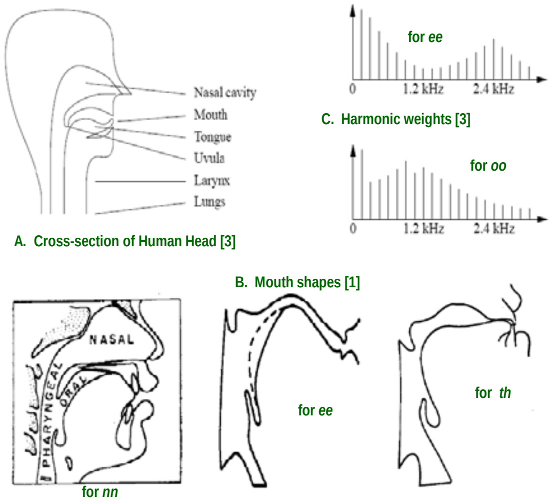

**Figure 1. Anatomy and Spectrum of Speech [1,3].**

The taxonomy of English phonemes is tabulated.

| Type | unvoiced | voiced |
|---|---|---|
| *Vowel-like* | | |
| mouth only | - | vowels, ll, rr |
| + nose | - | mm, nn, ng |
| diphthongs | - | ow, long-i, … |
| *Fricatives* | hh | wh |
| (sustained | ss | zz |
| turbulence) | sh | zh |
| | ff | vv |
| *Plosives* | ch | j |
| (burst | k | g |
| turbulence) | p | b |
| | t | d |

*Sustained* phonemes include vowels, nasals, *ll*, *rr*, and *fricatives. Dynamic* phonemes include *diphthongs* and *plosives.* The last eight rows in the table represent eight different mouth positions. We produce two phonemes from each position by vibrating our larynx, or not.

The acoustic signal's spectrum runs from $f_1$ to our hearing limit of 14-20 kHz, depending on the listener's age, etc. The acoustic energy in different phonemes is distributed differently over the aural spectrum; for example, fricatives like *ss* have significant energy at the high end of the spectrum. Hearing accuracy is a non-linear function of how much of this spectrum is actually heard.

### Review of Human Hearing

In each ear, the drum is *AC-coupled* (the Eustachian tube maintains DC) to the *cochlea* by tiny bones. See Figure 2. Shaped like a snail-shell, the cochlea is filled with fluid and lined internally with small hairs. The acoustic signal causes fluid waves inside the cochlea to excite nerves at the base of each hair.

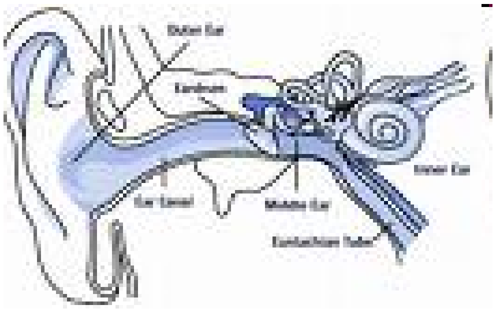

**Figure 2. Anatomy of Hearing [wikipedia].**

These nerves transmit a parallel signal to the brain giving the weights of the signal's harmonics. It appears that the cochlea and its *driver* in *brainware* compute the Fourier-Series coefficients of the received acoustic signal. Behind this aural *driver,* mid-level brainware performs more processing; it:

1. Calculates acoustic directionality,
2. Selects the desired signal out of background noise and other intelligible signals,
3. Performs phoneme discrimination (independent of the speaker),
4. Identifies who the speaker is (independent of the phoneme).

The last three tasks are supported by high-level syntactic and semantic processing which, at even higher levels of brainware, depend on content, context, background, and emotional state. While it's all quite remarkable, this paper deals only with low- and mid-level brainware and the last two tasks on the list above.

**Review of Aural Processing**

It's believed that a mid-level brainware process *identifies speakers* by comparing the set of weights, received from the driver, against a speaker database. Our accuracy at finding a best match is a nonlinear function of how many weights the *speaker-identifier process* receives from the driver. The number of coefficients depends on how much acoustic spectrum is heard by the cochlea and its driver.

It's believed we *discriminate phonemes* more indirecttly. The spectral envelope of most phonemes has four relative maxima, called *formant frequencies* $F_1$ to $F_4$. $F_1$ and $F_2$ peaks for *ee* and *oo* are apparent in the frequency domain in Figure 1C. Figure 3 shows generalized time-domain diagrams of $F_1$ and $F_2$ for nine phoneme pairs, each a dynamic consonant that elides into a vowel.

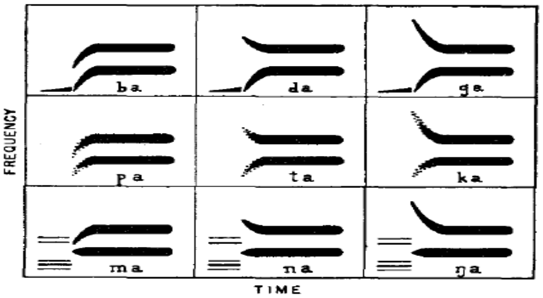

**Figure 3. [$F_1$,$F_2$] for nine phoneme pairs [1]**

The position on the spectrum of these formants, especially $F_1$ and $F_2$, seems to be the most important cue in phoneme discrimination. But, it's complex because formant positions are speaker dependent. Each point in Figure 4 is the [$F_1$, $F_2$] value as 76 people speak ten sustained phonemes. The clusters show the intended phoneme and the proximities indicate the error potential with no added spectral information. For example, the upper-left cluster represents *ee.* Low $F_1$ and high $F_2$ are consistent with *ee*'s spectrum on Figure 1B. We see a high potential that *ee* might be interpreted as *short-i.*

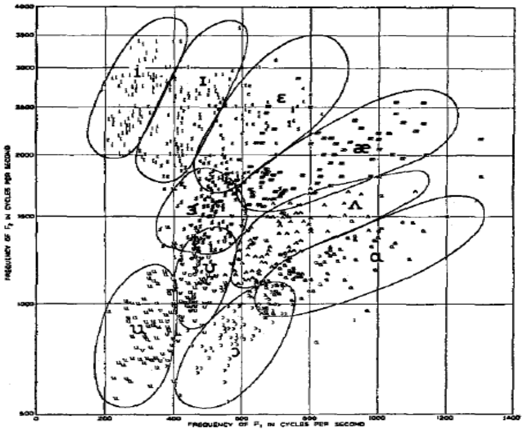

**Figure 4. [$F_1$,$F_2$] of 10 phonemes & 76 speakers [1].**

It's believed that we *discriminate phonemes* in a mid-level brainware process that computes the formants from the set of weights received from the driver and compares them against a phoneme database that works like Figure 4. Our accuracy at finding the best match is a non-linear function of how many formants the *phoneme discriminator process* has available to it. This number of formants depends, again, on how much of the acoustic spectrum is heard by the cochlea and its driver.

It's thought we have a *mirrored set* of multilevel processes in the speaker's brainware also. The communicating processes translate thoughts into language, and then to the sequence of neural signals that control our mouth parts.

## 3. TECHNOLOGY'S IMPACT ON QUALITY

After listing components of aural quality, this section reviews successive technologies and how they raised some aspects of audio quality and lowered others. After discussing their effect on speaker identification and phoneme discrimination, we review the history of the complaint that technology should never lower any aspect of application quality.

**Aural Quality and its Impairments**

Aural *quality* is measured by its *intensity, purity, immediacy, clarity* (small distortion), and *fidelity*. While *fidelity* really measures an audio signal's faithfulness to its acoustic analog, we'll defer to the lay use that it implies high bandwidth. Shown across the top of Figure 5, natural acoustic signals suffer five natural impairments: *loss, noise, crosstalk, delay,* and *echo.*

The role of networks is to eliminate natural *loss.* Also, they replace large acoustic *delay* by small signal delay and reduce other natural impairments listed second down on Figure 5's left. Analog networks add *crosstalk* from the loop pair and *echo* from impedance mismatch and leaky hybrids. And, they add new impairments not seen in natural signals: *amplitude distortion* from amplifiers that clip, *band-restriction* and *frequency distortion* from wire reactance, and *delay distortion* because different frequency components travel at different velocities [3].

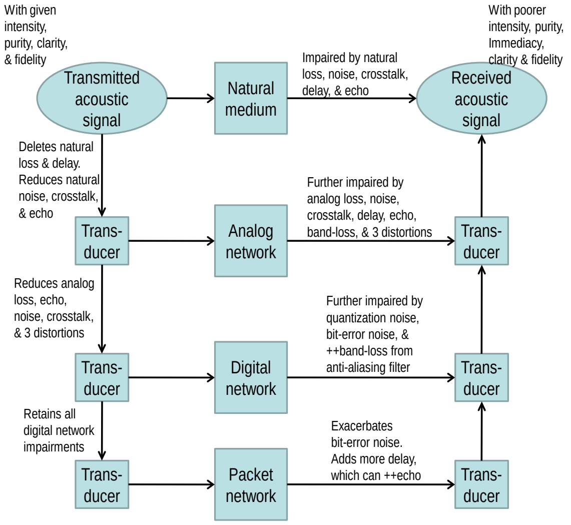
With given intensity, purity, clarity, & fidelity
Transmitted acoustic signal
Natural medium
Impaired by natural loss, noise, crosstalk, delay, & echo
With poorer intensity, purity, Immediacy, clarity & fidelity
Received acoustic signal
Deletes natural loss & delay. Reduces natural noise, crosstalk, & echo
Trans-ducer
Analog network
Further impaired by analog loss, noise, crosstalk, delay, echo, band-loss, & 3 distortions
Trans-ducer
Reduces analog loss, echo, noise, crosstalk, & 3 distortions
Trans-ducer
Digital network
Further impaired by quantization noise, bit-error noise, & ++band-loss from anti-aliasing filter
Trans-ducer
Retains all digital network impairments
Trans-ducer
Packet network
Exacerbates bit-error noise. Adds more delay, which can ++echo
Trans-ducer

**Figure 5. Layered Quality Impairments**

Analog networks degrade *fidelity* in three layers. As in Figure 6A, while 500-sets cut off $f_1$ at the low end, they had 12-kHz of bandpass. New telephones have less band-pass because modern networks give no reason to provide more. If phones are connected in a local call, as in 6B, the loop limits end-to-end bandpass to 8-10 kHz, depending on loop-length. In long-distance calls, as in 6C, the network further limits bandpass to 4-6 kHz, depending on distance. Though a 4-kHz analog long-distance channel had the worst *fidelity* of all connection types, the term *"toll grade"* was "spun" to imply high quality. Note that, the upper limit of all these bandwidths is given as a 3-dB frequency, but there is significant audio power above these formal limits.

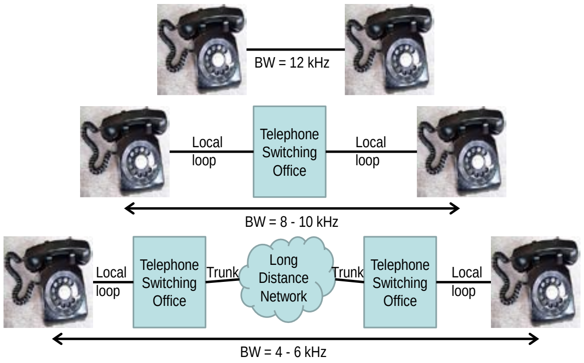

**Figure 6. Layered Reduction of Fidelity**

Analog technology advancements in channels (fiber), amplifiers, echo cancellers, shielding, and noise filters improved *loss, noise, crosstalk, delay, echo,* and *amplitude distortion* – but not *band-restriction* nor the other two forms of *distortion.*

Shown across Figure 5's third row, digitizing an audio signal improves *intensity*. *Quality* is higher because a digital PSTN is virtually noise-free (except for rare bit-errors) and the loop's noise (assuming the ADC is in the CO) is partially blocked on the speaker side by the ADC's anti-alias filter. But, new *noise* is added by quantizing, companding, mu-to-A conversion, and bit errors. *Echo* is worse because digital transport is usually four-wire, which requires many more hybrids (which can leak) in the network.

But, the worst impairment from digitizing voice is that anti-aliasing filters in the A-to-D converters impair *fidelity* such that all signals are nominally as band-limited as worst-case (long distance) analog signals. *Fidelity* is perceptibly even lower than nominal because blocking all audio above 4 kHz requires a half-power point at 3.7 kHz and high-end drop-off that is much steeper than in analog networks. So, while digital calls have much better SNR than analog calls, a local digital call has perceptibly lower *fidelity* than a long-distance analog call.

Seen across Figure 5's bottom row, VoIP adds a new layer of impairments to the quality of digital audio signals. Audio *purity* is further impaired because speech decompressors exaggerate bit errors and decoders cause noticeable *clunks* if packets are lost. Also, some silence-detecting codecs have a slow start-up that clips leading plosives.

*Immediacy* is greatly impaired by *delays* caused by packetization, jitter buffers, router processing, and multi-hop packet retransmission. VoIP calls often exceed user acceptance of conversation interaction delay. The user opinions tabulated below are a compromise between the Bell System's rigid standards and the IETF/ITU spin [2].

| Round-Trip Delay | Opinion |
|---|---|
| < 150 ms | good |
| 150-300 ms | noticeable |
| 300-450 ms | annoying |
| > 450 ms | unacceptable |

*Echo* worsens indirectly because human sensitivity to echo is delay-dependent [2]. Figure 7 shows how user complaints about echo vary with echo-to-signal ratio (TELR) and one-way delay. Since a digital conversation's TELR is about 55 dB, we see that round-trip delay should be less than 250-400 ms; but often it is not. So, VoIP-to-POTS and VoIP-to-cell (especially) calls are characterized by annoying echo.

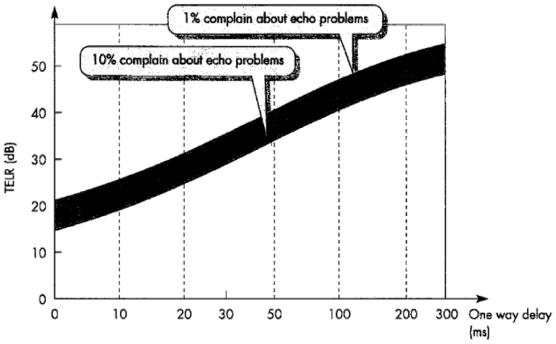

**Figure 7. User Annoyance with Echo [2]**

We see that, while digitizing speech offers a net overall improvement to audio quality, VoIP makes no positive contribution; it only lowers the quality. The last section proposes how we might change this.

### Identifying Phonemes and Speakers

“Telephone voice” impairs our ability to hear what a speaker says and identify who the speaker is. Since the 4-kHz DS0 channel has enough bandwidth for $F_1$ and $F_2$, we have little trouble identifying sustained non-fricative phonemes. Hearing the third and fourth formants might improve our discrimination of these sounds but, while the third may pass over a DS0 channel, the fourth typically will not.

We need a 7-kHz channel to receive all four formants and more than 7 kHz to better discriminate sounds we typically struggle with: nasals (distinguishing *mm* and *nn),* plosives (distinguishing *k* and *t)* and fricatives (distinguishing *ss* and *ff).* In one experiment [1], *ff* was spoken to many listeners over three channels with the following results:

| | Identified as: | | | |
|---|---|---|---|---|
| Chan BW | *ff* | *th* | *p* | *other* |
| 200-2500 Hz | 186 | 31 | 6 | 13 |
| 200-5000 Hz | 194 | 35 | 6 | 9 |
| 1000-5000 Hz | 162 | 28 | 12 | 50 |

It’s generally agreed that we identify speakers directly by their Fourier weights and not their formants. If so, our success would be based on the amount of data – the number of weights received. For three population groups, the table below shows the typical range of their fundamental pitch, the corresponding number of harmonics that would pass through a 4-kHz channel, and the group’s rank.

| Type | f1-range | # harmonics | Rank |
|---|---|---|---|
| Men | 75-150 Hz | 25-50 | most |
| Women | 140-300 Hz | 11-27 | middle |
| Children | 275-350 Hz | 9-13 | least |

This table is consistent with most people’s experience at speaker identification over the telephone: men are easy to recognize, women less easy, and we see why “all children sound the same on the phone.” It’s also clear that a child could be recognized over a 12-kHz channel as well as an average male is over a 4-kHz channel. At 12 kHz, women would be more identifiable than men at 4 kHz, and men would be almost perfectly identified.

### The History of the Complaint

This is not a new complaint. When T1 was proposed in the 1960s, Amos Joel objected to its 8-kHz sample rate. T1’s advocates stifled him by saying he was a *dinosaur* who objected to digital voice (he did not). Now, some VoIP advocates use this tactic to stifle their critics. 8-kHz sampling was standardized when bandwidth was expensive; now that it isn’t, we’re still stuck with the DS0 channel (or are we?).

## 4. NETWORK INTEGRATION AND APP-QUALITY

After reviewing historical attempts at integrating networks [5], this section proposes a generalization of how integration naturally lowers app quality [5] and asks why we have refused to learn this lesson.

### History of Integrated Networks

Thirty-five years ago, ISDN was proposed as a global end-to-end network for all data types. Today, it’s relegated to the network edge, as an access standard. ISDN’s post mortem shows two reasons it failed:

1. ISDN needed a global digital network, an inexpensive users’ appliance/terminal, and a collection of integrated services – simultaneously. AT&T could have done this but was too focused on surviving (it didn’t).
2. We learned that the application matters. Ethernet’s statmuxing was more efficient for bursty data, especially keystrokes on a LAN, than ISDN circuit switching. And, efficiency trumped integration.

Twenty years ago, we proposed ATM as a global end-to-end network to carry all data types. Using cell relay and virtual circuits, ATM cleverly avoided network congestion from large packets. While ATM had limited success in the network core, where packet congestion is significant, it failed to achieve its main goal. ATM’s post mortem shows two reasons:

1. ATM’s success required that it also be cost-effective as a LAN. But, Ethernet prevailed because of its embedded base of interface cards, LAN manager familiarity, and its evolution to higher rates
2. We saw again that application matters. ATM was compared to a duck: “Ducks can swim, fly, and walk, but none well. ATM carries voice, data, and video, but none well.”

Now, the Internet is proposed as a global end-to-end network to carry all data types. ISDN and ATM each failed in part because application matters. What is different now?

### Why Integration Lowers App Quality

We suggest an economic explanation using Figure 8. Consider four cases defined by separated/integrated networks and low/high app-quality.

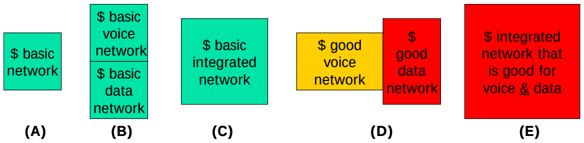

**Figure 8. Illustrating Network Costs [5].**

*1. Separated & low.* Let Figure 8A represent the cost of a basic un-optimized network. Suppose we have two apps, voice and data, with equal load. Then, 8B represents the cost of two separate networks, each dedicated to one app. App quality is barely acceptable because neither network has been optimized for its app’s quality.

*2. Integrated & low.* We could provide both apps over one (un-optimized) integrated network. The square in 8C, representing this integrated network, has larger area than the square in 8A because 8C supports twice as much load. But, 8C’s area is less than the sum of the areas of both squares in 8B because of economy-of-scale and a reduced staff of network managers. Since apps may interact in the integrated network, each app’s quality is slightly worse than in the pair of separate networks. This is the classic “duck”.

*3. Separated & high.* Let’s raise the quality of both apps by optimizing each network in 8B, which raises their cost. Figure 8D represents the cost of separate optimized networks. Each square in 8B is elongated to a rectangle in 8D, along a different dimension to indicate that each network is optimazed differently for its respective app.

*4. Integrated & high.* To improve the apps’ quality, or regain pre-integration un-optimized quality, we perform the same optimizations that were performed on the separate networks. So, the square in 8C, representing the “duck,” is elongated, but in in both dimensions. This gives the large square in 8E, which we call a "SWAN" (Superior-service-With-all-Apps Network).

If we don't care about app-quality, Case 2 beats Case 1 – the integrated network is slightly more economical (but not by very much). If we do care about quality, comparing Cases 3 and 4 isn't as easy. It's not clear how the area of 8E (representing the cost of a SWAN) compares against the sum of the areas of the rectangles in 8D.

Does the cost of optimizing an integrated network, so its apps have good quality, cancel out the small savings provided by the integration? Apparently so, or wouldn't IP-based voice carriers – Qwest long-distance, Skype, and Vonage – have dominated the telephone industry by now?

### Why are we Blind to this Lesson?

While the prior analysis is admittedly weak, it's not fundamentally flawed. It seems clear from this analysis, and the history lesson preceding it, that network integration is a bad idea (assuming we don't want to further degrade app-quality). Half a millennium ago, alchemists had a goal that is at least easy to appreciate. Our determination to continue trying to integrate networks is admirable, but puzzling.

## 5. WHAT CAN WE DO?

Ranting about how bad things are has become an all-too-familiar form of discourse. In an effort to more than rant, and make a positive contribution, this section makes the transition from how-bad-it-is to how-good-it-could-be by discussing the market potential and proposing a solution.

### Market Potential for High-Quality Apps

Before starting, we must ask if a significant market niche exists that cares about voice quality. Casually observing young people, we see many who have a taste for music that doesn't benefit from large spectrum and a high tolerance for the poor audio quality of cell-phones. If a market exists, it would be among people who appreciate the kind of music that does sound noticeably better over a high-fidelity channel and who are annoyed by, or even have difficulty with, the audio quality of their cell-phones. This second group tends to be older than the first, and it is growing rapidly as the surge of baby boomers become older and deafer. Note that their decreasing ear-bandwidth reinforces the adequacy of the 12-kHz channel.

The prior paragraph is not based on an accurate marketing study. But, it seems likely that, if the market size that justifies product development isn't significant enough yet, it may become large enough in just a few more years.

### High-fidelity Voice-over-IP

Some readers may think this article has bashed VoIP. It didn't. VoIP presents the opportunity to raise voice quality, not just to *toll-grade,* but even beyond.

As discussed, we can significantly improve phoneme discrimination and speaker identification using a 12-kHz channel. Since this bandwidth is triple the DS0's equivalent bandwidth, it may be accomplish easily by installing three DS0 codecs in an IP-phone, each taking 8000 samples per second, but 125/3 = 41.7 μs out-of-phase. This technique should also work with speech compressing codecs.

These three DS0 streams could be packetized together easily at the speaker's end and separated at the listener end. Downward compatibility is accomplished simply, by ignoring 2/3 of the data. While this proposal needs to be tested, two others have already been implemented and tested in the Telecom Program at the University of Pittsburgh.

1. VoIP delay, and echo's dependence on delay, can be reduced by optimal packetization [6,7]. When a network is lightly loaded, packetization delay is reduced by generating small packets often, perhaps every 10 ms. When a network is heavily loaded, network queuing delays can be reduced by generating larger packets less often, perhaps every 30 ms. This technique has been demonstrated and the signaling needed to synchronize end-points was implemented using VoIP's RTC Protocol [6,7].
2. The ITU defines overall audio quality as a complicated function of codec type, end-to-end delay, fidelity, and other issues [4]. IP-phones with multiple codec-types can optimize overall audio quality by changing codec-type midstream depending on network congestion [8,9]. Control signaling can also use VoIP's RTC Protocol.

At Pitt, we are in the process of building a prototype system, we call *Ernestine,* in which such techniques will be implemented and tested.

## 6. CONCLUSION

While technology has certainly improved audio quality over the last 100 years, some aspects of audio quality, especially fidelity, have devolved. But, this devolution has an ironic solution. VoIP's poor audio quality isn't inherent to VoIP, but is a function of design choices, some of which date back to the 1960s. Surprisingly, VoIP gives us an opportunity to provide excellent audio quality – if design changes proposed herein are implemented.

## REFERENCES

1. I Lehiste (editor), *Readings in Acoustic Phonetics,* MIT Press, 1967.
2. O Hersent, D Gurle, and JP Petit, *IP Telephony,* Addison-Wesley, 1999.
3. RA Thompson, *Telephone Switching Systems,* Artech House, 2000.
4. ITU-T G.107, *The E-model, a computational model for use in transmission planning,* 2000.
5. R Thompson, *The Ugly Duckling,* tutorial at IEEE's High-Speed Switching and Routing Conference, Oct. 2006 (slides available from the author).
6. RA Thompson, *Optimal Packetization to Minimize VoIP's End-to-End Delay,* ITERA Conference, Oct. 2006 (slides available from the author).
7. B Ngamwanwattana, *Optimizing Packetization for Minimal End-to-end Delay in VoIP Networks,* PhD dissertation, University of Pittsburgh, 2007.
8. E Myokhatnykh, *Adaptive Speech Quality in Voice-over-IP Communications,* PhD dissertation, University of Pittsburgh, 2007.
9. RA Thompson, *Controlling Speech Quality in VoIP Networks,* Presented to the Pittsburgh Section of the IEEE, May 2008 (slides available from the author).