会話音声の明瞭度に対する帯域の影響

ホワイトペーパー

Jeff Rodman
特別研究員 兼 最高技術責任者
2003 年 1 月 16 日
jrodman@polycom.com

要約

テレフォニの会話音声の明瞭度に影響する要素のうち、帯域は最も重要な要素の 1 つであることが明らかになっています。このホワイトペーパーでは、帯域と会話音声の明瞭度との関連性を明らかにするとともに、帯域の拡張によって他の問題 (ノイズや残響など) をどのように補えるか、最新の音声通信システムにおいて音声用帯域を拡張することの重要性がどのように認識されているかを考察します。

はじめに

大陸を横断する最初の通話が 1915 年に実現して以来、問題の原因と解決策がさまざまな科学技術によって徐々に明らかになり、電話技術の欠点が緩和されてきました。なかでも音響学、物理学、化学、電子工学は、電話機の設計に大きな進歩をもたらし、送話口と受話器の設計にしても、1940 年までに 10 dB の周波数向上が達成されました。(1) また、これらのパーツのゲインをより正確に調整するための改良も行われました。初期の実験段階では、中の炭素粉が固まらないように、話し手がカーボンマイクを叩いたり揺すったりしなければなりませんでした。その後、電話が進化するにつれて防側音回路も追加され、話し手が自分の声の大きさを判断しやすくなりました。そして長距離電話が日常的になり、相手側のエコーが邪魔に感じられるようになると、回線にエコー抑制が追加され、後にデジタル エコー キャンセル機能も追加されました。

しかしながら、電話網で使用できる音声用の帯域に関しては、この 60 年間ほとんど進歩していません。初期の電話接続は意図的に制限されていたのではなく、当時のトランスデューサや機材の特性によって制限されていました。明瞭度の研究は一般に 4 kHz～8 kHz か、場合によってはそれを超える周波数で行われていましたが、1930 年代の電話網で伝達可能な信号は最大でも 3 kHz に過ぎず、最初のマルチチャネル通信システムに関しても、最大約 3.5 kHz でした。標準化においても、G.711 のデジタルテレフォニに関する規定においても、現在の電話網の周波数は、最大約 3.3 kHz として一般に容認されています。分割前の AT&T のベル研究所で 1984 年に行われた PSTN テストでは、短～中距離の接続において 3.2 kHz で著しいロールオフが見られ、長距離接続になると 2.7 kHz まで落ち込みました。(2) 電話網では、スペクトルの下限にあたる 220 Hz 未満の周波数を伝達することはできません。伝達可能な最低限度は、通常 280～300 Hz 程度です。

このような電話のパフォーマンスとは異なり、FM ラジオとテレビでは 30 Hz～15 kHz、CD オーディオでは 20 Hz～20 kHz、プロやオーディオマニアが使う高度なオーディオシステムでは 20 Hz～22 kHz 以上、そして AM ラジオでは最大 5 kHz を網羅しています。ポリコムの新しい VTX テクノロジでは、ビジネス電話の音声を標準アナログ電話回線上で 7 kHz まで引き上げることが可能です。また、IP システムのデスクトップ電話においても 7 kHz での運用が実現しつつあります。

帯域と明瞭度

1917 年、Crandall は「多くの単語は母音に注意しなくても文脈で判断できる。明瞭度の決定要素は子音である」と指摘しています。(3) 「take him to the map (彼を地図まで連れて行く)」と「take him to the mat (彼をマットまで連れて行く)」では、まったく意味が異なります。修理業者が「soffet (軒)」の修繕依頼を「faucet (蛇口)」と聞き間違えたために無駄な時間を費やすこともあるかもしれません。pole (柱)、bole (木の幹)、coal (石炭)、dole (施し物)、foal (子馬)、goal (ゴール)、told (言った)、hole (穴)、molt (脱皮)、mold (鋳型)、noel (クリスマス)、bold (大胆な)、yo (やあ)、roll (ロール)、colt (初心者)、sole (唯一の)、dolt (まぬけ)、sold (売約済み)、toll (電話料金)、bolt (ボルト)、vole (野ネズミ)、gold (ゴールド)、shoal (浅瀬)、troll (流し釣り) はすべて、母音が同じであり、組み合わされている子音が異なるだけですが、子音の相違によって文の意味がまったく異なって

きます。このように、子音はフランス語、ドイツ語、イタリア語、ポーランド語、ロシア語、日本語などの多くの言語で重要な役割を担っています。[4] もちろん、子音は会話の中で頻繁に出現します。たとえば、よく混同される p と t は、単純な会話に含まれる音素の 10 % を占めています。f と s で 6.8 %、m と n で 10.3 % になります。[5] 会話全体の音素の半分以上は子音なのです。

会話における子音の重要性は、電話網に対する深刻な課題となります。子音のエネルギーの大部分は高周波数帯で伝達され、ときには電話の帯域を完全に超えることすらあるからです。英語の場合、会話の平均的なエネルギーの大部分を占める母音は 3 kHz 未満の周波数帯にありますが、会話の最も重要な要素である子音はその上の周波数帯に存在します。たとえば、f と s の相違は 3 kHz を超える周波数帯に存在するため、電話の帯域である 3.3 kHz を超えてしまいます。図 1 で示すように、「sailing (航海)」の「s」を「failing (欠陥)」の「f」と区別するための高周波数音は 4 kHz～14 kHz の間に集中しています。この部分の周波数が欠落すると、相手が何を言ったのか理解する手がかりがなくなります。

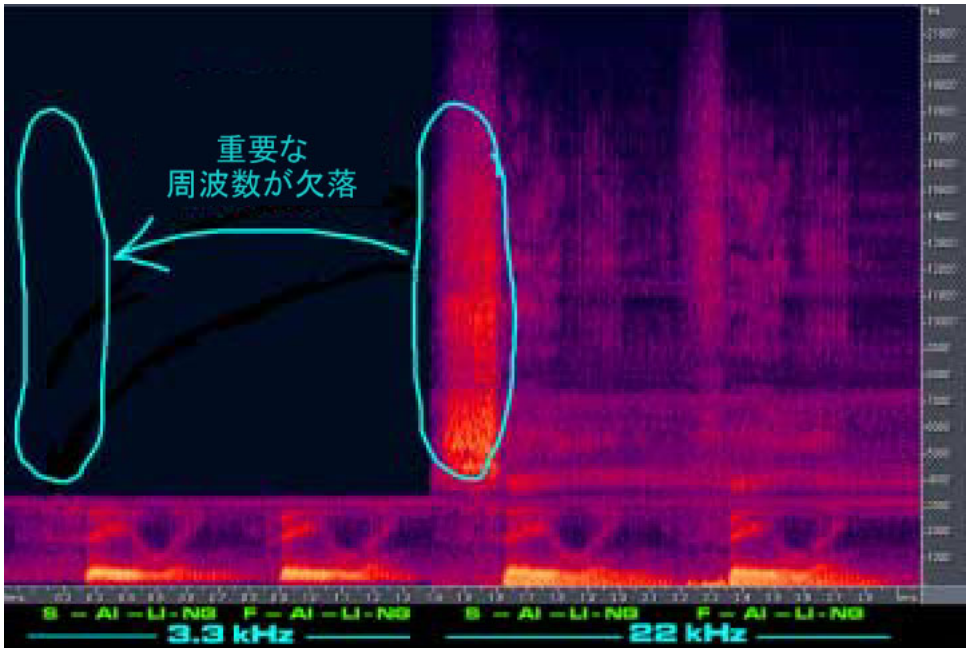

図 1: sailing と failing の音声スペクトルは 3.3 kHz および 22 kHz に分布

つまり、「my cousin is sailing in college (私のいとこは大学でヨット部に入っています)」と「my cousin is failing in college (私のいとこは大学で落第しそうになっています)」の違いを従来の電話で伝えるには、文脈 (たとえば、いとこがよくヨットの練習をしていると知っているなど) を加味することが不可欠だということです。

帯域、聴覚、発声力学

これまでを要約すると、人間の耳で最も感知しやすい周波数の 3 分の 2 と、会話時の周波数の 80 % が、公衆電話網の能力を超えていることになります。人間の耳は 3.3 kHz の音を最も感知しやすくできていますが、これは電話網の周波数帯がちょうど途切れるところです。知覚された音声の品質測定方法の標準である PAMS (知覚分析尺度システム) で評価すると、3.3 kHz の会話音声が 4 ポイントであるのに対し、4 7 kHz では満点の 5 ポイントになります。[6]

人間の声では、母音と子音が異なるメカニズムで形成されています。声帯の振動は、そのままでは低い "ブーン" という音です。この "ブーン" という音が声道を通って各周波数に形成される

ことで、母音が完成します。ちょうど、トランペットの奏者がマウスピースに吹き込んだ "ブー" という音が、管内で調整されるのと同じです。声道では、特定の周波数帯の音が各**フォルマント**に集められます。フォルマントとは、400 Hz、1200 Hz、2000 Hz の近辺で形成されるまとまりです。ただし、フォルマント周波数は最大 3900 Hz を超える場合もあり、その音声によって大きく異なります。(7)

一方、子音は無声のクリック音、破裂音、気息音などで構成されています。これらの音は、声帯からではなく、舌、頬、歯などを接触させたり、打ち鳴らしたり、歯擦音を出すことによって形成されます。母音や長い有声音の分析に使うと便利なフォルマントも、会話に含まれる情報の多くを伝達する要素である子音にはほとんど無関係です。

明瞭度、帯域、倦怠感

会話音声の精度を計測する場合、リストに書かれた音節、単語、文を聞き手のグループに読み聞かせるという方法が一般的です。ここでいう明瞭度とは、聞き手によって正しく記録された割合を意味します。明瞭度指数、単語明瞭度、音節明瞭度はすべて、話し手の言った単語を聞き手が正しく判断できる度合いを計測する尺度です。

計測の結果、会話音声の明瞭度は帯域の低下とともに低下することがわかります。単音節語の場合、帯域 3.3 kHz における精度が 75 % であるのに対し、帯域 7 kHz では 95 % になっています。(8)

このような明瞭度の低下は、文中で音が組み合わさることによってさらに悪化します。10 語から成る文の場合、各語の信頼性が 90 % であっても、明確に理解される可能性は 35 % (0.9^{10}) しかありません (図 2)。通常の会話スピードでは、1 分あたりの単語数は約 120 語です。したがって、3.3 kHz では毎分約 40 か所に曖昧性が生じることになります。これに対し 7 kHz の場合は 4 か所未満であり、電話回線を通さない生の会話に近い精度が得られています。

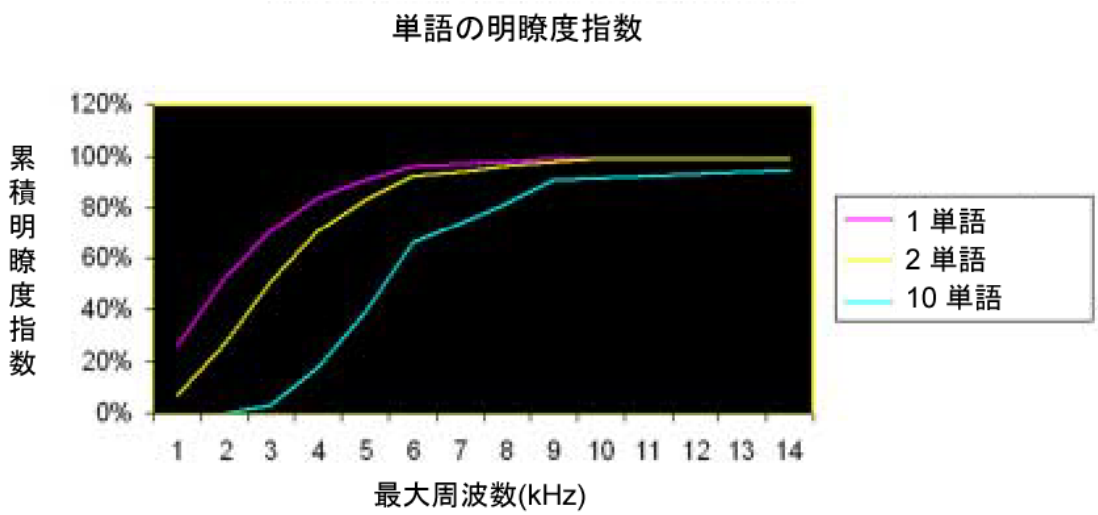

図 2: 低い帯域では文の明瞭度が低下

脳にはある程度の補完能力があるので、私たちはそれほど頻繁には混乱を自覚していません。ある音が明瞭でなければ、脳はその音の文脈を探ろうとします。最初に、「可能性のある文法のうちで何があてはまるか」という文法的な分析が行われます。たとえば、「I have to tie my choose (選択のひもを結ばなければならない)」では意味が通らないので、「choose」ではなく「shoes (靴)」だろうというように判断が下されます。

文法的に複数の可能性がある場合、聞き手は文脈から判断してどんな単語であれば意味が通るかを決定します。たとえば、海洋生物学者は「dolphin (いるか)」に関して述べているに違いない、フランス史学者が考えているのは「dauphin (王大使)」であろうと判断します。

しかし、会議が進行する中でこのような言葉のパズルが続くと、聞き手の集中力はそがれてしまいます。使われた単語が何であるかを推論しようとすると、意識がそれて会話の一部分を見失いないがちです。それが会議中に何度も起こると、倦怠感が高まり、理解度やインタラクションは低下します。聞き手は会話の流れについていくよりも、話し手の発した言葉が何だったかを理解する方に注意を払うようになります。聞き手の時間のあまりにも多くが、発言の内容を理解するよりも、正しい単語の究明に費やされてしまうのです。

音声の精度に影響するその他の要素

ビジネスで会議を実施する際には、音声帯域と相互に影響し合うその他の要素にも目を向ける必要があります。

残響

残響は、どんな室内でも生じうる自然な反響の一種ですが、帯域が限られている場合の音質劣化を増幅させる要素です。グループ間でのビデオ会議は会議室で行われるのが一般的ですが、会議室の残響時間は長いので、ビジネステレフォニを実施する上で大きな問題になります。また、話し手がマイクから遠ざかったときやマイクが話し手側を向いていないときにも、直接的な会話音声より残響の方が聞き手に伝わりやすいため状況が悪化します。

さらに、音声会議システムでは、聞き手が電話回線を通して使用できる "耳" (マイク) が 1 つしかないため、室内の残響が増幅されます。耳が 2 つあれば、残響の多くを脳内で除去することが可能ですが、このメカニズムは今日になっても解明されていません。しかし電話会議では信号を伝えるチャネルが 1 つしかないので、このような脳による残響除去機能がうまく機能しません。大きなホールで講演を聞く機会があったら、目を閉じて、一方の耳をふさいでみてください。

話の内容が理解しにくくなることに気づくはずです。話し手から 3 m 離れても明瞭に声が聞こえるのに、同じ場所に設置されたマイクを通した声を聞くと、それが高品質マイクであっても残響が強くぼんやりした感じになるのはこのためです。

帯域を広げることは、この問題の緩和に大きく役立ちます。あるテストでは、残響時間の長い室内で帯域を 4 kHz から 7 kHz に引き上げると、明瞭度が 52 % から 80 % に改善したと報告されています。[(9)]

言葉のなまり

拡大するグローバルビジネスの世界では、言語や方言の異なる人どうしが電話で話す場合の精度が重要になっています。なまりのある音声は、ネイティブの音声に比べて理解が困難です。これは、なまりそのものだけでなく、文法も発音も単語の選択までもが聞き手の予測と異なることにも起因しています。たとえば韓国人が英語で話す際には、f と p が置き換わることがよくあります (「faint (気絶)」は「paint (ペンキ)」、「coffee (コーヒー)」は「copy (コピー)」のように発音されます)。トルコ人は、単語に余分な音節を加えることがあります (「stone」は、「istone」や「sitone」のように発音されます)。話し手がロンドン出身である場合は、たばこ入れのことを「fag packet」と言うので (子音に注意)、聞き手がアメリカ人であればすっかり当惑するでしょう。

このようなことを考えると、音声通信における物理的なパラメータ (帯域、残響、増幅、インタラクション、ノイズ) がいかに重要であるかは明らかです。不明瞭な単語は文法的な文脈から推測可能である、という楽観的な意見はもはや通りません。そのような方法は、共通した正しい文法が使われているという仮定に基づいたものです。話し手と聞き手の生育環境が異なる場合、こ

の仮定が怪しくなります。そのため、話し手になまりがある場合は、音声帯域を拡大して精度を向上させることが一層重要です。

ささやき声

高周波数の相対的な重要性に影響するもう 1 つの可変要素は、ささやき声による会話です。長時間の会話を平均すると、通常の声の場合は 7 kHz のエネルギーが 600 Hz のエネルギーより約 40 dB も低くなりますが、ささやき声の場合はほとんど変化がなく、この 3 オクターブで 10 dB しか低下しません。そのため電話の帯域では、ささやき声で話すと母音さえもかなり不明瞭になります。風邪で喉を痛めている人が話し手である場合、電話の帯域では、声のしゃがれ具合に比例してエネルギーが小さくなることと、大きな声を出せないことから、さらに声が聞き取りにくくなります。

電話周波数の下限より低い周波数

会話に使用される音声スペクトルの一部は、電話周波数の下限 (300 Hz) を下回る低周波数帯にあります。話声に含まれる母音の基本周波数は、声帯が実際に振動する 100 Hz 付近に集中しています。図 3 では、母音 "o" のスペクトルを示しています。これを見ると、電話帯域の範囲 (黄色い帯) には、話声に含まれるエネルギーのほんの一部しか存在しないことが明らかです。

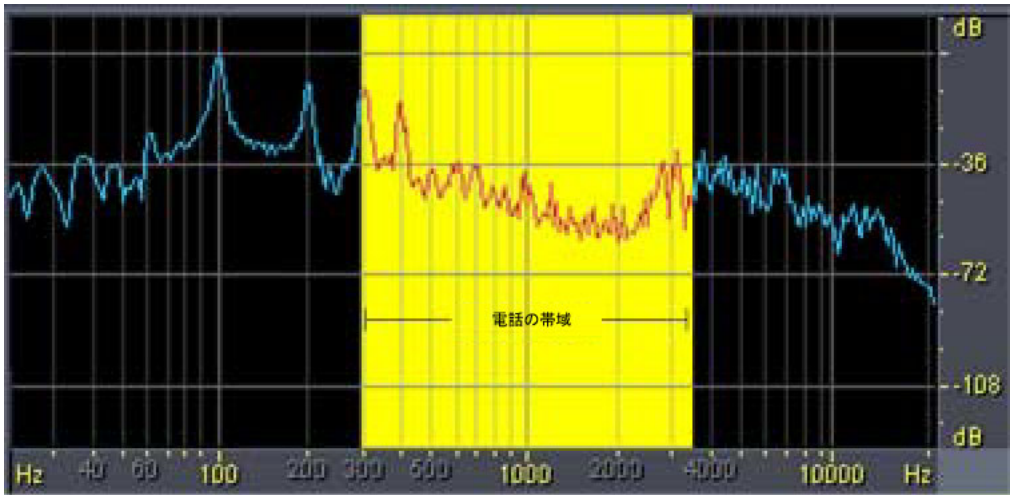

図 3: 母音 "o" (男声)

電話で伝達可能な帯域の上限または下限を超えた部分にも重要な周波数帯がありますが、電話ではそれが排除されていることがわかります。高周波数帯だけでなく、低周波数帯も重要です。電話での声が "非現実的" で快適性に欠け、どこか話し手の存在感を実感しにくいことが多いのは、250 Hz 未満の周波数帯がカットされるからです。これまでのテストでは、人の可聴域より低い超低周波域 (20～80 Hz) にも、p、b、k、t、d などの子音を聞き分ける重要な手がかりが含まれていることがわかっています。[(10)]

最新の音声通信システム

音声通信における帯域の制限によって生じる問題は、今日では広く認識されています。これを受けて、拡張帯域を利用できる音声通信システムが普及してきました。たとえばビデオ会議では、通常の音声接続に 7 kHz (場合によっては 14 kHz) が使用されています。FM ラジオとテレビでは、15 kHz 帯域の音声が伝達されています。IP テレフォニでも、TIA 920-200 に示されている圧縮/非圧縮コーデック技術を採り入れて、7 kHz 帯域へと移行しつつあります。携帯電話網においても、G.722.2 音声コーデックにより、音声周波数帯として 7 kHz へ拡張され始めています。

結論

電話の帯域を 7 kHz 以上に拡張すると、会議中の倦怠感、集中力、明瞭度を大幅に改善できることは明白です。このことは、実際の会議室の状況を改善する上で特に重要です。実際の会議室では、残響、プロジェクタやエアコンのノイズ、話し手のなまりなど、ビジネステレフォニに伴うさまざまな音響問題が発生し、音質が劣化することがよくあるからです。さらに、電話の帯域を 300 Hz 以下まで拡張すれば、存在感や現実感の問題を大きく改善できます。

1938 年、AT&T の Inglis は、電話システムの帯域について考察した論文で、「周波数の制限は本質的に経済的な制限であり、条件の変化によって変化しうる」と述べています。[11] 21 世紀になった今、Inglis が予見していたように経済も条件も変化しました。現代の電話技術では、帯域を拡張して明瞭な音声を実現することができるのです。

---

[1] A. H. Inglis 著、「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、K. I. Park 共著、「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 著、「The Composition of Speech」、Phys. Rev. 10 ser. 2 (1917): 75

[4] John Collard 著、「A Theoretical Study of the Articulation and Intelligibility of a Telephone Circuit」、Electrical Communication 7 (1929): 174

[5] P. B. Denes 著、「On the Statistics of Spoken English」、The Journal Of the Acoustical Society of America 35 (6) (1963): 892-904

[6] Anthony Rix および Mike Hollier 共著、「Perceptual speech quality assessment from narrowband telephony to wideband audio」、AES 107th Convention、New York: 24-27 September 1999

[7] Gordon E. Peterson 著、「The Information-Bearing Elements of Speech」、The Acoustical Society of America Journal、24 (6) (1952): 632

[8] N. R. French および J. C. Steinberg 共著、「Factors Governing the Intelligibility of Speech Sounds」、The Acoustical Society of America Journal、19 (1) (1947): 90

[9] P. W. Barnett 著、「Overview of Speech Intelligibility」、Proceedings of the Institute of Acoustics、21 Part 5 (1999)

[10] L. L. Myasnikov、E. M. Miasnikova、M. Y. Pekel'nyi 共著、「Infrasonic Cues for the Automatic Recognition of Speech Sounds」、Soviet Physics – Acoustics、14 No. 4 (April-June, 1969): 522

[11] A. H. Inglis 著、「Transmission Features of the New Telephone Sets」、Bell System Technical Journal 17 (1938): 358-380