SUBJECTIVE RATINGS OF INSTANTANEOUS AND GRADUAL TRANSITIONS FROM NARROWBAND TO WIDEBAND ACTIVE SPEECH

Stephen D. Voran

Institute for Telecommunication Sciences
325 Broadway, Boulder, Colorado, USA
svoran@its.bldrdoc.gov

ABSTRACT

In advanced heterogeneous telecommunication networks, network resources can dynamically dictate the type of speech coding that is used. An increase in resources allows for lower coding distortion or it might also be used to provide wideband speech instead of narrowband speech. Existing studies have demonstrated that wideband speech is preferred to narrowband speech, but they have also demonstrated that an abrupt transition from narrowband to wideband is perceived as an impairment, even though it is a transition to a higher quality signal. We describe our recent work that resulted in subjective scores for abrupt and gradual transitions from narrowband to wideband at the midpoint of a six-second segment of active speech. On average, signals that start narrowband and end wideband are rated slightly lower than constant narrowband signals and results are nearly the same for abrupt and gradual (2.5 second) transitions. Scores from 20 listeners show a wide range of individual opinions so we conclude that studies of bandwidth transitions may be quite sensitive to the listener population sample.

Index Terms— Narrowband speech, speech coding, subjective testing, wideband speech

1. BACKGROUND AND MOTIVATION

Telecommunication networks are simultaneously becoming less homogeneous and more adaptive. This makes it more likely that the network resources available to support a given call can change during the call, especially if one or more call participants are mobile. If network resources increase and additional data capacity is available then it is possible to switch to higher bit rate speech coding. One might keep the encoded speech bandwidth fixed and use the extra bits to reduce coding distortion. Another possibility is to switch from narrowband (NB) to wideband (WB) speech coding.

Unless otherwise indicated, we use what we consider to be the canonical definitions of speech passbands, based on the original NB and WB digital speech coders. Thus NB indicates a speech passband of 300 to 3400 Hz consistent with the minimum -3 dB bandwidth for G.711 PCM specified in [1] and WB refers to a speech passband of 50 to 7000 Hz as given in [2].

Several studies, including [3], have shown that WB is preferred to NB in controlled subjective experiments. Note that a transition from NB to WB entails both a low frequency extension (LFE) and a high frequency extension (HFE). The study in [3] shows that the HFE alone does not enhance perceived speech quality but the LFE alone does. If the LFE is in place then the HFE further enhances speech quality.

Consistent with this finding, more recent NB/WB speech coders include most or all of the LFE in the NB mode, and switching to WB entails mainly HFE. The specifications for the G.729.1 speech coding algorithm indicate a nominal NB bandwidth of 50 to 4000 Hz and a nominal WB bandwidth of 50 to 7000 Hz [4]. Measurements of the AMR speech coders show NB -3 dB bandwidth from 85 Hz up to 2800 to 3600 Hz (the upper limit depends on AMR mode) [5], and WB bandwidth from 50 Hz up to 5700 to 6600 Hz (the upper limit depends on mode) [6].

Given that listeners prefer WB to NB, it might seem logical to have a telecommunication system switch from NB to WB speech coding as soon as network resources become available. Earlier work shows that if listeners hear quality Q_{low} for a total of $(1-\alpha) \times T$ seconds and quality Q_{high} for a total of $\alpha \times T$ seconds, then as α goes from 0 to 1 the overall rating of the experience increases monotonically from Q_{low} to Q_{high} [7]. From this result we might expect a signal with both NB and WB portions to receive a quality rating between the constant NB rating and the constant WB rating, and thus there would be at least some *improvement* associated with switching to WB whenever possible. We note however that results in [7] were based on 3 second signals and at this short time-scale listeners were not conscious of the *transitions* between quality levels. In addition, this work did not use any bandwidth transitions.

More recently, a team of researchers developing and evaluating handoff strategies for telecommunications over wireless networks has included NB/WB switching in a set of important experiments [8] [9]. This is a rich body of work and it has revealed much about speech quality associated with handoffs, packet loss, bandwidth switching, and relationships among those factors. Here we focus only on the bandwidth switching aspect of this work.

Included in [8] and [9] are subjective tests with signals that switch between an NB speech coder (G.711) and a WB speech coder (AMR-WB also named G.722.2) in conjunction with a network handoff. Results indicate that the handoffs themselves do not hurt perceived speech quality but the bandwidth switching can hurt perceived speech quality. Specifically, switching from NB to WB coding at the midpoint of a six second recording results in a *lower* score (mean opinion score or MOS near 3.4) than a constant NB version of the recording (MOS near 3.9). WB speech coding was rated to have MOS near 5.0.

In 60 second tests, results show that switching from NB to WB near the 15 or 30 second point results in a small score increase relative to constant NB, but switching near the 45 second mark results in a small decrease. The conclusions are that *the switch from NB coding to WB coding is perceived as an impairment*, even though it is a transition to a higher quality speech signal. If this impairment happens early enough in the signal, it can be outweighed by the higher quality of WB in the remainder of the signal.

The notion that bandwidth switching is at least a minor impair-

ment can also be found in [10] which reports "In our experiments, switching between bandwidths, as long as it does not occur very frequently, is not distracting to listeners." But this observation seems to be focused on waveform continuity and the clicks or pops that may be created in the transition. These can easily be eliminated by simple waveform smoothing techniques.

In [6] we also find mention of infrequent bandwidth switching: "Although it is not expected that such switching appears on a frame-by-frame basis, it can happen e.g. once per call because of handover ..." It is not clear if this wording is intended to convey an expectation based on observations or to recommend system design goal.

The work in [6] , [8], and [9] used instantaneous transitions that had no pops or clicks. Would more gradual NB-to-WB transitions be less annoying to listeners? Note that when the G.729.1 speech coder switches from NB to WB the HFE is faded in over a period of one second [4]. Our investigation of this question progressed in three stages and these are described in the next three sections of this paper. The final stage was a subjective experiment that evaluated instantaneous and gradual (up to 2.5 second long) NB-to-WB transitions using uncoded NB and WB speech signals, six seconds long, with a transition starting at the three second point. Signals with transitions were rated slightly lower than constant NB signals, and results were nearly the same for abrupt and gradual transitions. Analysis of 20 listeners shows a wide range of individual preferences and we conclude that studies of bandwidth transitions may be quite sensitive to the listener population sample.

2. AUDIBILITY OF LOW AND HIGH FREQUENCY EXTENSIONS

We seek to design transitions between NB and WB speech that gradually fade in the LFE and HFE over a time window. This can be implemented without redundant transmission as shown in Figure 1. At sample number $k = k_0$ both switches flip up from the NB path to the WB path. But $g(0) = g_{min}$ is small so the WB signal is forced through the bandpass filter, returning it to an NB signal and preventing an abrupt bandwidth change. As time progresses $g(k - k_0)$ increases and this incrementally bypasses the filter. When $g(k - k_0) = 1$, the full WB signal is delivered.

The initial value of the fade-in g_{min} should be below the threshold of audibility to prevent the perception of the extensions "turning on." But any portion of the fade-in that is below the audible threshold increases the duration of the fade-in without adding any value. Thus we should set g_{min} safely below, yet near, the threshold.

Figure 2 shows a conceptual block diagram for determining thresholds. All filters are linear phase FIR with order 512 (lowpass) or 1024 (highpass). Values in the figure are -3 dB points. Passband ripple is less than 0.02 dB, and stop-band response is attenuated by more than 55 dB. Transition bands are 70 Hz (lowpass) or 35 Hz (highpass) wide.

The signal $y(k)$ is given by

$$y(k) = x_{NB}(k) + g_L \cdot x_L(k) + g_H \cdot x_H(k)\,, \qquad (1)$$

so $y(k)$ can be manipulated between NB and WB. Setting $g_L = g_H = 0$ gives an NB signal, while $g_L = g_H = 1$ gives a WB signal. A stereo sound card was used to play x_{NB} and y simultaneously and a passive switch allowed listeners to switch between level-normalized versions of x_{NB} and y at will. The normalization forced the two signals to have the same average A-weighted power.

The sound card (Echo Audio Mia) has frequency response that is flat to within $+0.1$, -1.0 dB across the band from 50 Hz to 7 kHz. This card feeds a headphone amplifier (Crown D-75A) with specified response of $\pm$ 0.1 dB from 20 Hz to 20 kHz. Circumaural headphones (Sennheiser HD 600) were used as the listening instrument and these have a specified -3 dB bandwidth from 16 Hz to 30 kHz.

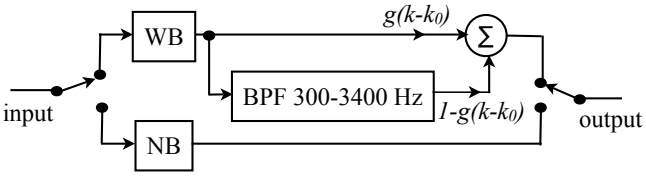

Fig. 1. Implementation of gradual bandwidth-switching transition.

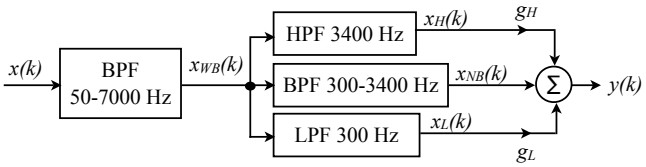

Fig. 2. Generation of signals for LFE and HFE detection tests.

Listeners used the switch to compare two unlabeled signals (x_{NB} and y) and indicated when the two signals sounded different. This protocol, combined with slow increases in g_H and/or g_L led to thresholds of audibility for the LFE and HFE for three female and three male listeners, using Harvard phonetically-balanced sentences from female and male talkers at preferred listening level in a sound isolated room (background noise measured below 20 dBA SPL).

For the LFE alone the median threshold of audibility is $G_L = 20\log_{10}(g_L) = -24$ dB ($g_H = 0$). For the HFE alone the median threshold of audibility is $G_H = 20\log_{10}(g_H) = -14$ dB ($g_L = 0$). These results indicate that the HFE alone is harder to detect than the LFE alone, at least for this small group of listeners. This provides an interesting parallel to the fact that the HFE alone is not perceived as an improvement, but the LFE alone is [3] . When both the LFE and HFE are presented together with the same gain, the median threshold of audibility is $G_L = G_H = -25$ dB with a range of -31 to -20 dB.

We expect that these thresholds form lower bounds for those that would be found in real telecommunications environments. In real environments background noise, limited transducer frequency response, the inability to make instantaneous "A vs. B" comparisons, and divided attention will likely raise the thresholds. Thus we selected $G_{min} = 20\log_{10}(g_{min}) = -30$ dB as a safe value for the start of the fade-in.

3. SEEKING AN OPTIMAL TRANSITION LENGTH

We have determined that $G(k) = 20\log_{10}(g(k))$ should increase from $G_{min} = -30$ dB up to 0 dB over some time period. Increasing the perceived loudness at a constant rate seems an intuitive choice for minimizing the audibility of this change. Decibels can serve as a rough surrogate for loudness, so we elected to increase $G(k)$ at a constant rate and this matches the choice made in [4]. When the fade-in starts with sample k_0, the duration of the transition is τ, and the sample rate is f_s we have

$$g(k) = \min\left\{10^{(G_{min}/20)(1-(k-k_0)/(\tau f_s))}, 1\right\}, \quad k_0 \le k. \quad (2)$$

It remains to pick τ and we theorize that this could be a trade-off. Larger values of τ might create transitions that are harder to detect and less annoying, but they would delay the arrival of the full WB signal. Smaller values might create transitions that are easier to detect and more annoying, but they would allow the full WB signal to begin sooner.

We used both the parameter optimization algorithm given in [11] and manual techniques to search for a τ value that would maximize the perceived speech quality of six-second signals with an NB-to-WB transition at the midway point. After six listeners provided perplexing results and some illuminating comments, we elected to investigate a small set of τ values in the next step. With hindsight it seems that we were trying to maximize a function that is essentially flat.

4. SUBJECTIVE SCORES FOR NARROWBAND, WIDEBAND, AND TRANSITIONS

Next we developed a subjective experiment using the MOS scale [12] to investigate four different NB-to-WB transitions. We digitally extracted phrases from spoken word CDs and converted the sample rate to 16,000 samples/second. The resulting recordings were each between 5.5 and 6.5 seconds in length, contained only active speech (no significant pauses), and covered four female and four male English talkers, each one saying two phrases (16 distinct phrases total). We processed these 16 recordings through 7 conditions as described in Table 1.

Condition	Description
1	WB
2	NB
3	NB $\rightarrow$ WB, $G_{min} = -30$ dB, $\tau = 2.5$ s
4	NB $\rightarrow$ WB, $G_{min} = -30$ dB, $\tau = 300$ ms
5	NB $\rightarrow$ WB, $G_{min} = -90$ dB, $\tau = 900$ ms [4]
6	NB $\rightarrow$ WB, $G_{min} = -30$ dB, $\tau = 0.3$ ms
7	NB MNRU, $Q = 15$ dB SNR

Table 1. Experiment conditions.

We used no speech coding, only bandpass filtering (with the specifications described in Section 2) to obtain results that speak to speech bandwidth alone and are not confounded by any speech coding distortions or other speech coding issues. The first two conditions provide NB and WB references and the final condition provides a low quality NB reference point via the modulated noise reference unit (MNRU). Conditions 3-6 were produced by transitioning from NB to WB according to Figure 1 and (2) with A-weighted power normalization added to prevent loudness shifts. In every case the transition started at the three second point in the recording.

Condition 3 used $\tau = 2.5$ seconds. This is the longest transition that allows some (i.e., 0.5 sec) WB signal at the end of the recording. Condition 6 used $\tau = 0.3$ ms which is just long enough to remove any waveform discontinuities. Condition 5 is the transition specified in [4]. This transition is prescribed to do nothing for the first 100 ms, then to fade in the extension(s) from -90 dB to 0 dB over the next 900 ms. Condition 4 is the portion of that transition that we expect would actually be audible, based on results in Section 2.

The laboratory and equipment were as described in Section 2. Twenty listeners (12 female and 8 male) participated. Their estimated ages ranged from 12 to 60 with a median near 40. All were unfamiliar with the experiment. Each listener first participated in a short practice session that included conditions 1, 2, 3, and 7. These results were discarded. Then the listener was presented with 88 sequentially numbered trials in a single self-paced session that lasted 12 to 15 minutes. After each trial the listener was instructed: "Please select your overall impression of the entire six-second recording." Replays were not allowed. This single session was actually built from two subsessions. The first (trials 1-32) included conditions 1, 2, 3, and 7, each crossed with a phrase from each of the 8 talkers and presented in a different random order to each listener. The second subsession (trials 33-88) used all seven conditions, each crossed with the other phrase from each of the 8 talkers and presented in a different random order to each listener.

The goal of this subsession structure was to obtain scores for the gradual transition of condition 3 (as well as the references 1,2, and 7) both before and after the listener was exposed to the more abrupt transitions of conditions 4-6. None of the condition means differed significantly (at the 95% level) between the two subsessions, so they have been combined in the results presented here. The consequence is that conditions 1, 2, 3, and 7 each have $20 \times 8 \times 2 = 320$ ratings while conditions 4-6 each have $20 \times 8 \times 1 = 160$ ratings.

The condition means and their 95% confidence intervals are shown in Figure 3. In this experiment switching from NB to WB at the midpoint of a six-second active speech recording incurs a slight penalty (MOS near 3.35), compared to constant NB (MOS near 3.55). This penalty is barely significant at the 95% level. This result is slightly different from the corresponding result in [9]. In [9] the transition conditions were scored about 0.5 MOS units below NB, in the present experiment the penalty is only about 0.2 MOS units. Note that both experiments found an average MOS difference between WB and NB near 1.1 MOS units.

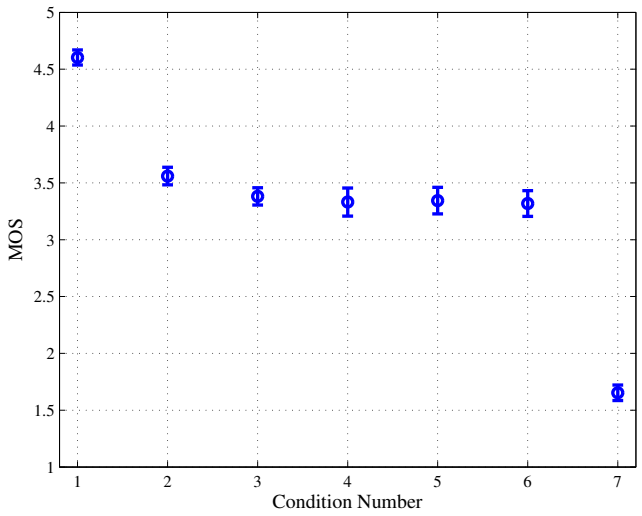

Fig. 3. Means and 95% confidence intervals for 7 conditions.

The experiment showed no significant difference between the average scores for the four different NB-to-WB transitions. In this experiment transition time is intrinsically coupled with the duration of the WB signal heard after the transition. It is possible that these two factors cancel each other out (e.g., abrupt annoying transition but 3 seconds of WB vs. gradual, less annoying transition but only 0.5 seconds of WB). But it is also possible that the different transitions are simply equally annoying, on average. Analysis of the individ-

ual listener results for conditions 3-6 seems to argue for this second possibility: only 1 of the 20 listeners showed any statistically significant (95% level) preference among the 4 transition conditions (that listener rated condition 3 superior to conditions 4-6). Intuition may suggest that gradual bandwidth transitions would be more palatable than abrupt ones, but no such preference was observed within the context of this experiment.

Finally, we consider the responses of the individual listeners. In light of the previous discussion, we average all transition conditions for each listener to create a per-listener transition score Q_t. We also calculate per-listener NB and WB scores Q_{NB} and Q_{WB}. Figure 4 shows the value of $Q_t - Q_{NB}$ for each listener, plotted as a function of $Q_{WB} - Q_{NB}$. Thus the horizontal position of any given point shows how much that listener values WB over NB, while the vertical position shows how that listener rates the transition conditions relative to NB. We have used four marker types to highlight four classes of listeners.

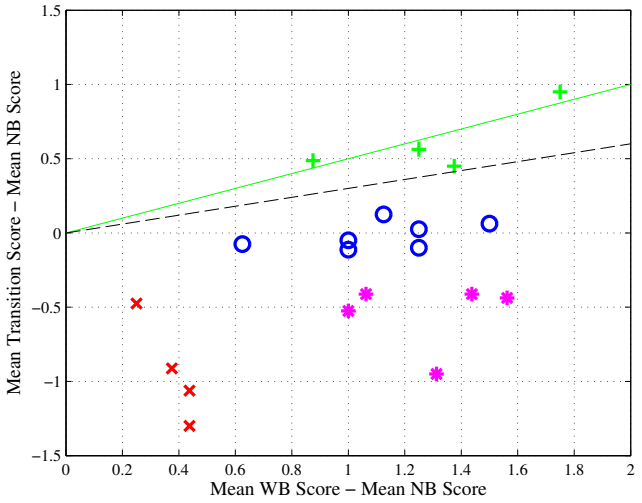

Fig. 4. Individual listener results, see text for details.

The circles show a group of 7 listeners who had a range of opinions on WB, but seem unmoved by the NB-to-WB transition. That is, a transition from NB to WB did not rate much higher or lower than constant NB. The asterisks indicate a group of five listeners who had relatively high opinions of WB, but found the transitions to be annoying, resulting in Q_t values from 0.4 to 1.0 below Q_{NB}. Four "x" symbols denote a group of listeners who had relatively low opinions of WB and were also annoyed by transitions, resulting in Q_t values dropping as low as 1.3 below Q_{NB}.

The final group also contains four listeners and is marked by plus signs. These listeners had a range of relatively high opinions of WB and they combined their NB and WB opinions to produce Q_t somewhat greater than Q_{NB}. The dashed line indicates simple averaging, $Q_t = (Q_{NB} + Q_{WB})/2$, and three of the four listeners seem to be using this rule. The fourth listener may be following the rule given in [7] shown by the dotted line. We did not detect any gender or age trends associated with these four groups.

Figure 4 shows that individual listeners considered the transition conditions to be worse than, the same as, or better than constant NB and (with a few possible exceptions) this behavior is not directly tied to the listeners' opinions of WB. Discussion of condition averages alone obscures this important diversity of opinions.

5. CONCLUSION AND DISCUSSION

We have found that an NB-to-WB transition in the middle of a six-second active speech recording is rated slightly lower, on average, than constant NB speech and this holds for all four transitions tested, independent of the transition duration. But behind this average result lies a wide range of individual opinions.

It may be that the transition to WB cannot be advantageous on average unless the WB signal continues for a longer period after the transition as in [8]. Or it may be that longer transition times or optimized transition gain functions $g(k-k_0)$, possibly individualized for the LFE and the HFE could lead to more pleasing transitions. Alternately, NB might be augmented with artificial bandwidth extension to mitigate the transition to true WB [6]. Finally, the use of delays and hysteresis might prevent unnecessary short-duration bandwidth changes that some listeners would certainly find annoying.

We extend warm thanks to Blazej Lewcio for helpful correspondences that motivated and set the stage for the work reported here.

6. REFERENCES

[1] ITU-T Rec. G.712, "Transmission performance characteristics of pulse code modulation channels," Geneva, 2001.

[2] ITU-T Rec. G.722, "7 kHz audio-coding within 64 kbit/s," Geneva, 1988.

[3] S. Voran, "Listener ratings of speech passbands," in *Proc. of the 1997 IEEE Workshop on Speech Coding for Telecommunications*, September 1997, pp. 81–82.

[4] ITU-T Rec. G.729.1, "G.729 based embedded variable bit-rate coder: an 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729," Geneva, 2006.

[5] "Performance characterization of the adaptive multi-rate (amr) speech codec," Tech. Rep. TR 126 975, ETSI, January 2009.

[6] "Performance characterization of the adaptive multi-rate wideband (amr-wb) speech codec," Tech. Rep. TR 126 976, ETSI, January 2009.

[7] S. Voran, "A basic experiment on time-varying speech quality," in *Proc. of the 4th International MESAQIN (Measurement of Speech and Audio Quality in Networks) Conference*, Prague, Czech Republic, June 2005, pp. 51–64.

[8] S. Moller, M. Waltermann, B. Lewcio, N. Kirschnick, and P. Vidales, "Speech quality while roaming in next generation networks," in *Proc. IEEE International Conference on Communications, ICC '09*, June 2009, pp. 1–5.

[9] B. Lewcio, S. Waltermann, M. Moller, and P. Vidales, "E-model supported switching between narrowband and wideband speech quality," in *Proc. of the First International Workshop on Quality of Multimedia Experience, QoMEX 2009*, July 2009, pp. 98–103.

[10] Bo Wei, Hui Dong, and J.D. Gibson, "Application of nb/wb amr speech codecs in the 30-khz tdma system," *IEEE Transactions on Wireless Communications*, vol. 3, no. 6, pp. 1897–1901, November 2004.

[11] S. Voran and A. Catellier, "Gradient ascent paired-comparison subjective quality testing," in *Proc. of the First International Workshop on Quality of Multimedia Experience, QoMEX 2009*, July 2009, pp. 133–138.

[12] ITU-T Rec. P.800, "Methods for subjective determination of transmission quality," Geneva, 1996.