A Basic Experiment on Time-Varying Speech Quality

Stephen D. Voran

Institute for Telecommunication Sciences, svoran@its.bldrdoc.gov

Abstract

We present a general formulation of a basic open question regarding the perception of time-varying speech quality. We then describe the design, implementation, conduct, and analysis of a practical experiment that addresses a small but fundamental part of that open question. In this experiment, listeners rate the overall speech quality of single sentence stimuli that contain two different levels of nominal speech quality and two transitions between these levels. We present several results including those related to human integration of speech quality and the recency effect. Finally, we discuss these results and suggest potential additional work that might build upon them.

Keywords

Robust speech coding, subjective listening experiment, time-varying speech quality

1 Background

The quality of speech delivered by a telecommunications network can depend on many factors. Some of these factors are usually constant for the duration of a call, but other factors can vary during a call. The result is a call that can contain time-varying speech quality. This adds significant complexity to the measurement of the speech quality delivered by the network, whether that measurement is subjective or objective. Lacking any other information, the conventional solution to this problem has been to simply calculate a time average and this is indeed an intuitive and simple starting place.

Several subjective experiments have been conducted in attempts to better understand various aspects of time-varying speech quality. Reference [1] describes an experiment where the Modulated Noise Reference Unit was used to impose one of two different speech quality vs. time profiles on 40-second speech recordings. Experiment subjects were asked to indicate their perceptions of speech quality in real-time using a sliding control. Subjects demonstrated the ability to do this task accurately, with a delay of about one second.

In [2] G.723.1 speech coding combined with packet losses was used to create five nominal quality levels. Ten different quality vs. time profiles were then created by switching between the five nominal quality levels. These 190-second recordings were presented to subjects who indicated their perceptions of speech quality in real-time using a sliding control. Each subject also gave an overall quality score at the conclusion of each 190-second recording. This type of work was expanded in [3] and [4] to consider conversation tests as well as listening tests, and real listening environments as well as laboratory listening environments.

In an experiment described in [5], short recordings (about 8 seconds each) were processed in various ways to create a range of quality levels. These short recordings were then combined in groups of seven to create longer recordings (about 1 minute). Subjects rated the speech quality of the short recordings in isolation, as well as the quality of the long recordings. In [6] six short recordings of various quality levels (about 5 seconds each) were combined to create longer recordings of about 30 seconds. Here also, subjects rated the speech quality of the short recordings in isolation, as well as the speech quality of the long recordings.

Finally, numerous experiments have been conducted on stochastic speech quality profiles. These experiments typically contain random or bursty occurrences of lower speech quality (called impairments) which commonly result from real or simulated packet loss in VoIP or fading in wireless telephony. Two examples can be found in [7] and [8]. The mean rate (and possibly other statistics) of the impairments are controlled, but the time history of speech quality is not controlled. The results of these experiments are typically relationships between the controlled statistics and overall perceived speech quality.

Each of the experiments in [1]-[6] have generally used different types of stimuli, different testing protocols, and different definitions of short-term and long-term (or overall) speech quality. Thus they are not directly comparable and their results do not directly reinforce or contradict each other. However, several results arise from one or more of these experiments that appear to be more general and seem likely to transcend the specific experiment designs.

Analysis of [5] and [6] reveals that long-term perceived speech quality scores are lower than the time average of the corresponding short-term perceived speech quality scores. These two experiments also reveal that the minimum of the short-term scores is a quantity that can be useful for estimating the long-term scores. These observations lead us to the notion that long-term speech quality may be some function of short-term speech quality that is bounded above by the mean function and bounded below by the minimum function.

In addition, analysis of [2] and [6] reveals that variation alone can reduce perceived long-term speech quality. That is, if the mean and minimum values of short-term qualities are held constant while the variance of short-term speech qualities increases, then the long-term speech quality will generally drop.

The recency effect is well established: long-term scores are more strongly influenced by events near the scoring time than by earlier events [2],[5].

In addition, time constants for human response to speech quality decreases are shorter than for speech-quality increases [2]. That is, listeners more quickly detect and/or report decreases in speech quality than increases in speech quality. We have also found that viewers of video signals share this trait.

2 Time-Varying Speech Quality Scenarios and Questions

There are numerous scenarios that lead to time-varying speech quality. One class of scenarios is associated with acoustic background noise conditions at the talker location that vary relative to the acoustic speech power that the talker or talkers provide. Another class is associated with impairments to the transmission channel between the talker and listener locations. These impairments are common in wireless telephony and VoIP systems. Researchers have characterized these impairments and have developed numerous robust coding techniques to mitigate their effects on the perceived quality of the delivered speech.

2.1 Robust Coding

Robust coding involves taking bits away from source coding and adding them to a channel coding scheme or using them in some other way to increase robustness. This creates a trade-off: removing more bits will lower the baseline speech quality, but it will increase robustness. An example is given in Figure 1. The upper panel shows a hypothetical example of a transmission channel quality history using arbitrary time and channel-quality units. The middle panel shows four different resulting hypothetical speech-quality histories, again using arbitrary units. Each of these shows an exaggerated "digital cliff effect:" speech quality is constant at some usable level for all channel qualities above some threshold and is constant at some low, unusable level for all channel qualities below that threshold. Line 1 represents a coding scheme with high baseline speech quality but very low robustness. Thus this high baseline speech quality can only be delivered when the channel quality is very high. Line 2 represents a more robust coding scheme, but with lower baseline speech quality. Line 3 represents an even more robust scheme and line 4 represents the most robust scheme. This final scheme has very low baseline speech quality, but it can deliver that speech quality reliably, even when the channel quality is very low.

Once a time and quality scale have been added to these speech-quality histories, there are many interesting, relevant questions that one might ask. For example: Which is associated with the higher overall perceived speech quality; the constant, low quality line marked 4, or the line marked 3 that has higher quality most of the time and a single period of failure? How does this result depend on the length of that failure period and the two baseline speech qualities? How would the results change with the inclusion of additional failures? In general, a better understanding of the perceived speech quality associated with speech-quality histories like those in the middle panel of Figure 1 can allow for better informed choices in the robust coding trade-off.

2.2 Adaptive Coding

If the time scale in Figure 1 is long enough and channel quality feedback is available, then it is possible to make intelligent adaptations to the coding scheme. These adaptations may seek to select the robust coding scheme that results in the highest short-term speech quality at any given time. Examples can be found in [9] and [10]. Ideally, this would result in the speech-quality history shown in the lower panel of Figure 1. A better understanding of the perceived speech quality associated with an arbitrary speech-quality history could aid such adaptation algorithms.

2.3 Multiple-Description Coding

A related, yet different, set of time-varying speech-quality issues are present when multiple-description coding (MDC) is used [11]-[15]. In the two-channel case of MDC, different descriptions (encodings) of the speech are sent on two different physical or virtual channels. If both channels successfully deliver a description, those two descriptions are combined to create an approximation to the original speech signal. (We call the associated speech quality the "two-channel speech quality.") If either channel fails to deliver its description, then the other description is used alone to generate a lower quality approximation to the original speech signal. (We call this the "one-channel speech quality," and this is typically constant regardless of which channel fails.)

One key parameter in a two-channel MDC system is the redundancy between the two descriptions. If the two descriptions are identical, they are completely redundant, and the two-channel speech quality is the same as the one-channel speech quality. If the two descriptions are completely independent, they have no redundancy, and the two-channel speech quality can be much higher than the one-channel speech quality. The true benefits of MDC are usually found between these two extremes.

Figure 2 shows an example set of speech-quality histories for MDC with three different levels of redundancy, with a fixed total bit rate and a fixed channel failure pattern. Line 1 shows the speech quality history for the case of minimally redundant descriptions, line 2 is for moderately redundant descriptions, and line 3 is for fully redundant descriptions. This figure leads to the question: "Which of these three options has the highest perceived overall speech quality?" More generally, when channel failure statistics are known, it is natural to ask which of the available options gives the highest overall perceived speech quality and to select the level of MDC redundancy accordingly. A better understanding of the perceived speech quality associated with an arbitrary speech-quality history could help to answer this question.

2.4 General Formulation of a Basic Open Question

An expanded and generalized version of the leftmost portion of the speech-quality history shown in the lower panel of Figure 1 is given in Figure 3. While this figure comes from the adaptive coding example above, with seven free parameters it can be viewed as a general and fundamental building block for almost any time-varying speech-quality history and it could be produced by any one of a large number of scenarios in robust, adaptive, or multiple-description speech coding and transmission.

The most general question motivated by Figure 3 is how the perceived overall speech quality relates to the seven variables shown in that figure. In other words, we would ideally seek a function $Q(q_L,q_H,t_1,t_2,t_3,\tau_1,\tau_2)$ that describes the perceived overall long-term speech quality as determined by the two quality variables and the five time variables. We will refer to this speech quality as the "overall speech quality." We will use the name "short-term speech quality" for the quantities q_L and q_H.

Given the observations in Section 1, we would expect that the function $Q(q_L,q_H,t_1,t_2,t_3,\tau_1,\tau_2)$ would be upper bounded by the time average and lower bounded by the minimum:

$$q_L \le Q(q_L,q_H,t_1,t_2,t_3,\tau_1,\tau_2) \le q_L + (q_H - q_L)\left(\frac{t_2 - t_1}{t_3}\right). \quad (1)$$

Within the limits of human perception, we intuitively expect $Q(q_L,q_H,t_1,t_2,t_3,\tau_1,\tau_2)$ to increase with q_L, q_H, and $(t_2 - t_1)$. Where the recency effect applies, we would also expect an increase with $\frac{1}{2}(t_1 + t_2)$, since this corresponds to moving the higher quality (q_H) segment of speech towards the end of the signal and thus towards the time when the speech quality judgment is made. Increasing τ_1 and/or τ_2 may make speech quality transitions harder to detect, and thus may reduce perceived speech quality variation and increase perceived overall speech quality. Beyond these general guidelines, the existing results do not provide further specific insight into the nature of $Q(q_L,q_H,t_1,t_2,t_3,\tau_1,\tau_2)$.

Complete knowledge of this function would be a basic yet fundamental step towards understanding perceived quality of time-varying speech in general. For example, it would increase our understanding of the perceived overall speech quality associated with an adaptive coding speech-quality history like the one shown in the lower panel of Figure 1 and it would allow an informed choice between the different MDC speech-quality histories shown in Figure 2. In addition, such information might be used in objective estimators of perceived speech quality to combine shorter-term speech-quality estimates to create longer-term estimates.

The function in question depends on seven variables. To properly explore this seven-dimensional space would require a huge family of experiments, and separate additional experiments would then be required to ascertain the generality of the results. Thus, as a starting place, we have designed a single, practical, experiment to gain some insight into a simplified and constrained version of the function $Q(q_L,q_H,t_1,t_2,t_3,\tau_1,\tau_2)$.

3 Experiment Description

The simplified and constrained case of time-varying speech quality addressed in this experiment is shown in Figure 4. A segment of higher speech quality is constrained to be temporally centered (at $\frac{t_3}{2}$ seconds) in the speech stimulus. In addition, the transitions in Figure 4 are shown as vertical lines because they are perceptually instantaneous. (The parameters τ_1 and τ_2 of Figure 3 are each zero.) Waveform discontinuities at these transitions are prevented by crossfades of 1 ms duration. Under these constraints, there are four, rather than seven, free parameters: $q_L, q_H, t_2 - t_1$, and t_3.

The experiment is an absolute category rating mean opinion score (MOS) experiment [16]. The response scale includes five options: excellent, good, fair, poor, and bad. Each stimulus in the experiment is an English-language sentence from the Harvard phonetically-balanced sentence lists [17]. Four female and four male talkers are used, and each provides five different sentences. All inactive speech portions are removed from the start and end of each recording and the resulting lengths range from t_3 =2.4 to 3.2 seconds. The mean and median lengths are both 2.7 seconds.

We have picked the burst duration factor

$$\alpha = \tfrac{1}{t_3}(t_2 - t_1) \quad (2)$$

as the single time-related control variable in the experiment and have selected seven discrete levels for that variable: $\alpha = 0.0, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0$. Note that when $\alpha = 0.0$, the entire stimulus is associated with the speech quality level q_L, and there is no switching in the stimulus. Similarly when $\alpha = 1.0$, the entire stimulus is associated with the speech quality level q_H.

In this experiment, q_L and q_H are selected from one of four discrete short-term quality levels. These four levels were created by 4-kHz nominal bandwidth speech coding in clear channel conditions. The four speech coding algorithms were selected to be relevant to telecommunications and to provide quality levels that are spread across the response scale to the extent possible. These four speech coding algorithms are: NATO MELP at 1.2 kbps [18], ETSI AMR-NB Mode 0 at 4.75 and Mode 5 at 7.95 kbps [19], and G.711 PCM at 64 kbps [20]. We refer to the associated nominal speech quality levels as $q_1, q_2, q_3,$ and q_4 respectively. These levels are "nominal" because speech quality can vary somewhat for a fixed speech coding algorithm due to variations in talkers, sentences, levels, background noises, etc.

There are a total of just three controlled variables in this experiment: q_L, q_H, and α. A "condition" is defined by a set of specific values for q_L, q_H, and α. The experiment contains 40 conditions and these are defined in the first five columns of Table 1. The first 35 conditions follow a common pattern and form 5 groups of 7 conditions per group, as indicated in column 2 of Table 1. In each of these five groups, q_L and q_H are held constant and α steps through the seven selected values in an increasing fashion. To keep the experiment size practical, these five groups cover five of the six possible combinations of q_L and q_H. These five combinations were selected for their coverage of the full range of possible values of q_L, q_H, and q_H - q_L.

The final five conditions (36-40) are different from the first 35 conditions, and were created to complement conditions 16-20, found in group 3. These conditions contain a burst of lower quality q_L centered in a stimulus that starts and ends with quality q_H as shown in Figure 5. For these conditions, $1-\alpha$ describes the duration of the burst of lower quality: $1-\alpha = \frac{1}{t_3}(t_2 - t_1)$. Thus αt_3 describes the total duration of the segments of higher quality, as is the case with the first 35 conditions.

Thus for i = 16 to 20, both condition i and condition i+20 have speech quality $q_H = q_4$ for αt_3 seconds, and speech quality $q_L = q_1$ for $(1-\alpha)t_3$ seconds. Condition i and condition i+20 differ in that condition i contains the higher level of speech quality (q_4) at the center of the stimulus, while condition i+20 contains the higher level of speech quality (q_4) at the start and end of the stimulus.

This set of five complementary conditions provides opportunities to detect any recency effects in this experiment and to test for perceptual equivalence between complementary speech-quality time histories. We refer to this set of conditions as group $\tilde{3}$. We define conditions 15 and 21 to be members of both group 3 and group $\tilde{3}$. Since the speech quality levels q_1 and q_4 are maximally separated, groups 3 and $\tilde{3}$ should provide maximal sensitivity to any recency effect. In addition, groups 3 and $\tilde{3}$ provide the two extreme temporal arrangements (q_4 at the center of the stimulus vs. the start and end of the stimulus) and this also should provide maximal sensitivity to any recency effect.

In general, perceived speech quality is influenced by talker, sentence, and speech coding algorithm. For each group of conditions, we are interested in how overall perceived speech quality changes with α. To minimize variance due to talker and sentence, we use a fixed set of talkers and sentences to evaluate each group. Thus, for each listener, α is the only experiment parameter that changes within each group of conditions. Also, the same set of talkers and sentences is used in group 3 and group $\tilde{3}$.

Each of the conditions in the experiment was evaluated by a total of 20 listeners, 11 female and 9 male. Listeners responded to the question "Please mark your opinion of the overall speech quality." The experiment was conducted in a sound-isolated room. Stimuli were played at preferred listening level through both earpieces of a high-quality headphone set. A total of 112 votes were collected for each condition and these were approximately balanced (female vs. male) from both the talker and listener perspectives. None of the listeners were involved in speech coding research and none were aware of the composition of the experiment stimuli.

4 Experiment Results

A numerical treatment of the 112 votes for each condition is enabled by equating the five vote categories (excellent, good, fair, poor, and bad) with the integers 5, 4, 3, 2, and 1 respectively [16]. For each condition, the resulting mean value (MOS) and the 95% confidence interval about that mean are given in columns 6 and 7 of Table 1. The half-widths of the 95% confidence intervals in this experiment range from 0.12 to 0.17, with a mean and median value of 0.15. We use the term "confidence interval" as shorthand for "95% confidence interval" throughout the remainder of this paper. The MOS values for each of the 40 conditions are also shown organized by group and as a function of α in Figure 6. This figure shows how perceived speech quality increases with the burst duration factor α.

4.1 Constant Speech Quality

Note that conditions 1, 8, and 15 each provide a MOS value for q_1, but via different talkers and sentences. This apparent redundancy is required by the constraints that each group of conditions be tested with a fixed set of talkers and sentences, yet the experiment had to contain multiple sets of talkers and sentences to prevent listener fatigue. Similarly, conditions 7 and 22 provide MOS values for q_2, conditions 14 and 29 provide MOS values for q_3, and conditions 21, 28, and 35 provide MOS values for q_4. For each of these groupings, the confidence intervals about the MOS values overlap, indicating consistent results in spite of the variation of talkers and sentences. These results can thus be combined to generate single MOS values and associated confidence intervals, based on 3×112 (q_1 and q_4) or 2×112 (q_2 and q_3) total votes. These results are given in Table 2. In this experiment, the four chosen nominal speech quality levels were perceived to span a range of about two points of the total four-point range.

4.2 Recency

Figure 6 allows for a graphical comparison of the groups 3 and $\tilde{3}$. If a recency effect were active in this experiment, we would expect group $\tilde{3}$ results to be consistently and significantly higher than group 3 results, since group $\tilde{3}$ stimuli end with higher speech quality while group 3 stimuli end with lower speech quality. A quick review of the MOS values and confidence intervals for conditions in these two groups makes it clear that this is not the case. This experiment reveals no reliable recency effect at these time scales.

4.3 Equivalence

In addition, an alternate view of groups 3 and $\tilde{3}$ allows us to conclude that stimuli with base quality q_L and bursts of higher quality, q_H, using burst duration factor α have the same perceived overall speech quality as stimuli with base quality q_H and bursts of lower quality, q_L, using burst duration factor $1-\alpha$. That is, when the appropriate, intuitive, equivalences are enforced, bursts of increased quality are equivalent to bursts of decreased quality. We have demonstrated this for the case of extreme speech qualities (q_1 and q_4) and thus fully expect it to extend to the more moderate cases where speech quality differences and transitions are harder to detect and are perceptually less significant.

4.4 Transition Locations

For each of the six groups shown in Figure 6, one could draw a monotonically increasing line that passes through each of the confidence intervals that surround each MOS value. When one looks at the MOS values in isolation, however, one sees several departures from the expected behavior. Specifically, the MOS values of conditions 23, 12, and 31 appear to be somewhat greater than a smooth monotonically increasing function of α would suggest. We believe that this may be a manifestation of a second-order effect related to the temporal locations of the speech-quality transitions relative to the syllabic structure of the stimuli.

We calculated the average power in the 20 ms windows centered at each transition for each condition. The three lowest values are associated with conditions 23, 12, and 31 which have average speech powers at the transitions that are 6.3, 5.4, and 4.9 dB respectively below the grand average transition speech power. Our hypothesis is that detectable transitions in speech quality detract from overall speech quality, and that transitions between syllables (lower average power) are harder to detect than transitions within syllables (higher average power).

Additional experiments directed specifically at this possible effect would be required to confirm or disprove this hypothesis, and to separate any such effects associated with upward transitions in speech quality from any effects associated with downward transitions in speech quality. Note that controlling both α and the locations of transitions relative to the speech structure would likely require searching through large amounts of speech to find suitable sentence-talker combinations. Note also that the notion of detectable transitions detracting from speech quality is consistent with the inverse relationship between short-term speech quality variation and overall speech quality presented in Section 1.

4.5 Time-Varying Speech Quality

The main effect in the experiment is the increase in perceived speech quality from q_L to q_H as α increases from 0 to 1. To study this effect, we transform all results into a normalized domain where the condition MOS values in each group increase from 0 to 1 as α increases from 0 to 1. For a given group of conditions, let $\mu(\alpha)$ represent the MOS for the condition with burst duration factor α. Then for that group of conditions, the normalization is accomplished by first subtracting $\mu(0)$ from all seven MOS values and then dividing the seven results by $\mu(1)-\mu(0)$.

In this normalized domain, for each fixed value of α, none of the 15 possible pairs of confidence intervals is disjoint. Thus we consider each group to display the same underlying, normalized, functional dependence on α, *independent of* q_L *and* q_H. Thus we average the normalized MOS values for the six groups at each value of α resulting in the curve shown in Figure 7. This figure also includes the straight line associated with temporal averaging

$$Q_A = q_L + (q_H - q_L)\alpha\,, \tag{3}$$

which becomes simply

$$Q_A = \alpha \tag{4}$$

in this normalized domain (since q_L and q_H correspond to zero and one respectively). Thus Figure 7 confirms that under the conditions of this experiment, the perceived speech quality is bounded above by the temporal average, and bounded below by the minimum value of speech quality q_L.

Analysis of the averaged normalized-domain results shown in Figure 7 leads us to two mathematical fits for the experimental results. The parabolic fit and the hyperbolic sine fit each have just one free parameter.

We optimize in the original domain (using unaveraged data) to minimize root mean-squared error (RMSE) between the fit and the MOS values generated by the experiment for each condition. These fits are

$$Q_p(q_L, q_H, \alpha) = q_L + (q_H - q_L)\left(a\alpha^2 + (1-a)\alpha\right) \tag{5}$$

and

$$Q_s(q_L, q_H, \alpha) = q_L + (q_H - q_L)\left(\frac{\sinh\left(\alpha/b\right)}{\sinh\left(1/b\right)}\right). \tag{6}$$

For this experiment, the optimizing values of the parameters are $a = 0.81$, and $b = 0.42$. For either fit, the resulting RMSE in the original domain is 0.08 MOS units. The parabolic fit passes through all but two of the confidence intervals about the experimental MOS values. The two exceptions are conditions 12 and 13. The hyperbolic sine fit passes through all of the confidence intervals except the one associated with condition 12. The unique behavior associated with condition 12 (and indirectly, its neighbor, condition 13) may be due to the locations of the transitions in this condition, as described in 4.4 above. Figure 8 through Figure 11 show the MOS values, confidence intervals, and the two fits. The six groups have been assigned to the four different figures for clarity of presentation.

4.6 Subconscious Integration

Informal exit interviews with some listeners revealed that they did not perceive two distinct levels of speech quality in the stimuli. That is, those listeners were not consciously combining multiple impressions before voting. Rather they were simply voting to report what they perceived as a single level of speech quality. This might be described as "subconscious integration" of speech quality.

It seems reasonable to hypothesize that "conscious integration" of speech quality could increase the variance in the votes, since listeners could exhibit individualized behaviors in this task. The experiment results show that the mean confidence intervals (and thus the mean variances) remain nearly constant (differing by only 0.02) between the class of conditions with nominally constant speech quality ($\alpha = 0,1$) and the class of conditions that include imposed speech quality variations ($0 < \alpha < 1$). Under the hypothesis that conscious integration could increase variance, this constant variance result is consistent with the notion of subconscious integration of speech quality.

5 Summary and Discussion

5.1 Summary

We have outlined previous experiments related to the perception of time-varying speech quality and have summarized those results that we believe are likely to be generally applicable, beyond the specific environment of any given experiment. We have described some of the sources of time-varying speech quality in telecommunications as well as the corresponding motivations for better understanding the perception of time-varying speech quality. In light of these preliminaries, we then presented a general formulation of a basic open question, followed by the design, conduct, and analysis of a single, practical experiment that addresses a small but fundamental part of that open question. The key results of this experiment are summarized again here:

1) Single sentence stimuli that take nominal perceived quality level q_H for αt_3 seconds and q_L for $(1-\alpha)t_3$ seconds have an overall perceived speech quality that is described equally well by (5) or (6). This is true within the context of this experiment, where each stimulus contains exactly two speech quality transitions and t_3 is on the order of three seconds.

2) Recency was not observed in the context of this experiment, even in the extreme cases that were selected to be most sensitive to recency. Thus result 1) above reveals a fundamental human integration operation that is not confounded by a speech quality history. Based on the discussion in 4.2 and 4.3, we expect these results to be valid for stimuli with segments of *higher or lower quality in any location*, subject to the constraints that exactly two transitions are present, and the stimulus duration is on the order of three seconds.

3) The location of a speech quality transition may affect its detectability and thus the overall perceived speech quality. As intuition suggests, transitions in syllables may be more detectable than those between syllables, and may result in lower overall perceived speech quality. Additional experimentation would be required to confirm this initial indication.

4) Based on listener reports, the speech-quality integration operation is subconscious at this time scale. Experiment variance results are consistent with this reporting.

5.2 Discussion

These results are far from completely general, but they do provide a starting place for this approach to the problem of time-varying speech quality. Potential follow-on experiments to ascertain how general or specific these results are could include the use of additional sources of distortion and different time scales. Other potential experiments to explore related topics could include experiments where the temporal locations of speech quality transitions are controlled in order to investigate how transition location affects transition detectability and perceived speech quality. Other experiments could investigate how the speech quality transition times (τ_1 and τ_2) relate to transition detectability and perceived speech quality.

Of these options, the investigation of longer time scales seems most important. The present results are based on stimuli with durations on the order of 3 seconds. Over what range of time scales will these results hold? How will they change outside that range? At what time scale does speech quality integration move from a subconscious task to a conscious task? Does variance increase when integration becomes conscious? In the context of typical telecommunications speech-quality variations, at what time scale does recency first occur? Is this in the regime of subconscious or conscious integration? When designing experiments for longer time scales, we suggest that it may be important to use a single, naturally connected stimulus, as one would experience in a typical telecommunications application. The concatenation of unconnected stimuli (e.g., different talkers, or different, unrelated sentences) may confound results by influencing listeners with an unintentional and undesired "timing signal" that is an artifact of the experiment.

Recency is a real effect for human subjects and the study of recency can reveal truths about human perception and behavior. If we wish to build objective estimators of perceived speech quality that emulate humans exactly, then we need to understand and emulate recency. If we take a step back and consider the ultimate purpose of objective estimators, the desirability of emulating recency may seem questionable.

In fact, we could argue that from a telecommunications engineering perspective, there are cases where recency is a nuisance. Consider an actual telecommunications service that delivers time-varying speech quality that is stationary in the statistical sense. When measuring and reporting overall perceptions of this service, it may be desired that those results be maximally representative of the general situation and thus maximally independent of the exact sampling time.

When listeners are used, and the time scale τ is long enough, recency is unavoidable. In this case it is likely that the only way to attain results that are independent of sampling time is to perform repeated measurement trials and average the results so that the recency effect will average out. This could then yield results that reflect human integration of speech quality over the interval τ without giving increased weight to the end of that interval. That is, long-term, recency-free, perceived speech-quality results.

The same averaging technique could be employed when using objective estimators. But emulating and then averaging out the recency effect is clearly inefficient from both the development and implementation perspectives. Instead, it may be desirable to directly emulate the long-term, recency-free perceived speech-quality results described above. This would require further experiments to determine the nature of human integration over longer time scales, with recency removed. The results developed in the present experiment are recency-free because of the short time scale. It is interesting to ponder whether or not these results might describe underlying human integration characteristics that are also operative at longer time scales, and that would be recognized once recency is removed.

6 References

[1] M. Hansen and B. Kollmeier, "Continuous Assessment of Time-Varying Speech Quality," *J. Acoust. Soc. Am.*, vol. 106, pp. 2888-2899, Nov. 1999.

[2] L. Gros and N. Chateau, "Instantaneous and Overall Judgements for Time-Varying Speech Quality: Assessments and Relationships," *Acta Acustica · Acustica*, vol. 87, pp. 367-377, 2001.

[3] L. Gros, "The Impact of Listening and Conversational Situations on Speech Perceived Quality for Time-Varying Impairments," in *Proc. International Conference on Measurement of Speech and Audio Quality in Networks*, pp. 17-19, Prague, Czech Republic, Feb. 2002.

[4] L. Gros, N. Chateau, and S. Busson, "Effects of Context on the Subjective Assessment of Time-Varying Speech Quality: Listening/Conversation, Laboratory/Real Environment," *Acta Acustica united with Acustica*, vol. 90, pp. 1037-1051, 2004.

[5] J. Rosenbluth, "Testing the Quality of Connections having Time Varying Impairments," Committee T1 Standards Contribution, T1A1.7/98-031, Oct. 1998.

[6] P. Gray, R. Massara, and M. Hollier, "An Experimental Investigation of the Accumulation of Perceived Error in Time-Varying Speech Distortions," Preprint, Audio Engineering Society 103rd Convention, New York, Sep. 1997.

[7] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder," *IEEE Trans. Speech and Audio Processing*, vol. 6, pp. 116-130, Mar. 1998.

[8] A. Lakaniemi, J. Rosti, and V.I. Raisanen, "Subjective VoIP Speech Quality Evaluation Based on Network Measurements," in *Proc. IEEE International Conference on Communications*, pp. 748 – 752, Helsinki, Finland, Jun. 2001.

[9] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Järvinen, "The Adaptive Multirate Wideband Speech Codec (AMR-WB)," *IEEE Trans. Speech and Audio Processing*, vol. 10, pp. 620-636, Nov. 2002.

[10] G. Yang, E. Shlomot, A. Benyassine, J. Thyssen, S. Huan-yu, and C. Murgia, "The SMV Algorithm Selected by TIA and 3GPP2 for CDMA Applications," in *Proc. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 709-712, Salt Lake City, USA, May 2001.

[11] A.A. El Gammal and T.M. Cover, "Achievable Rates for Multiple Descriptions," *IEEE Trans. Information Theory,* vol. IT-28, pp. 851-857, Nov. 1982.

[12] L. Ozarow, "On a Source Coding Problem with Two Channels and Three Receivers," *Bell System Technical J.*, vol. 59, pp. 1909-1921, Dec. 1980.

[13] R. Arean, J. Kovačević, and V.K. Goyal, "Multiple Description Perceptual Audio Coding with Correlating Transforms," *IEEE Trans. Speech and Audio Processing*, vol. 8, pp. 140-145, Mar. 2000.

[14] S. Voran, "The Channel-Optimized Multiple-Description Scalar Quantizer," in *Proc. 10th IEEE Digital Signal Processing Workshop*, Pine Mountain, Georgia, USA, Oct. 2002.

[15] S. Voran, "A Multiple-Description PCM Speech Coder using Structured Dual Vector Quantizers," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, Mar. 2005.

[16] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality," Geneva, 1996.

[17] "IEEE Recommended Practice for Speech Quality Measurements," *IEEE Trans. Audio and Electroacoustics*, vol. AU-17, pp. 225-246, Sep. 1969.

[18] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. Collura, "A 1200/2400 bps Coding Suite Based on MELP," in *Proc. 2002 IEEE Speech Coding Workshop*, Ibaraki, Japan, Oct. 2002.

[19] K. Järvinen, "Standardization of the Adaptive Multi-Rate Codec," in *Proc. European Signal Processing Conference*, Tampere, Finland, Sep. 2000.

[20] ITU-T Recommendation G.711, "Pulse Code Modulation (PCM) of Voice Frequencies," Geneva, 1988.

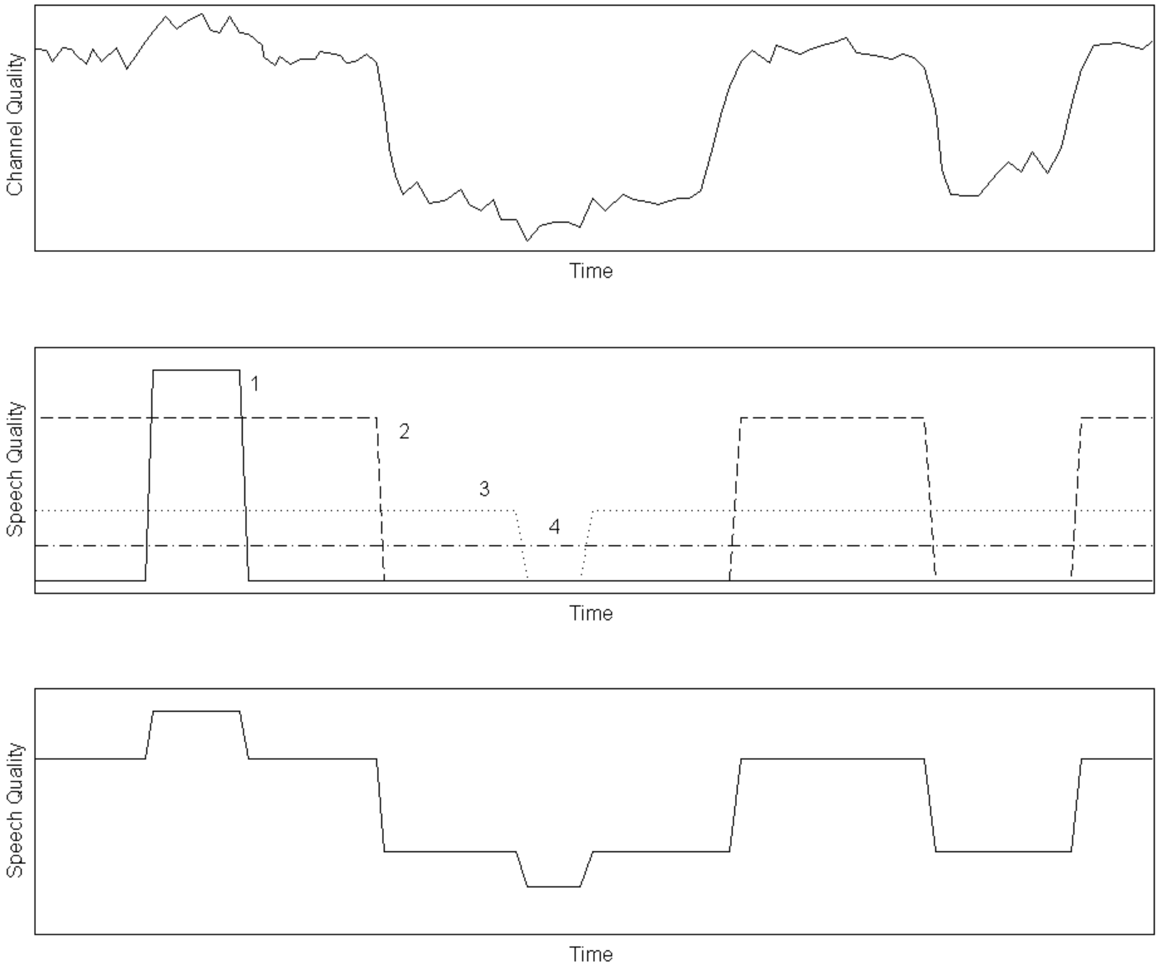

Fig. 1. Hypothetical example of speech quality vs. robustness trade-offs and the adaptive coding solution.

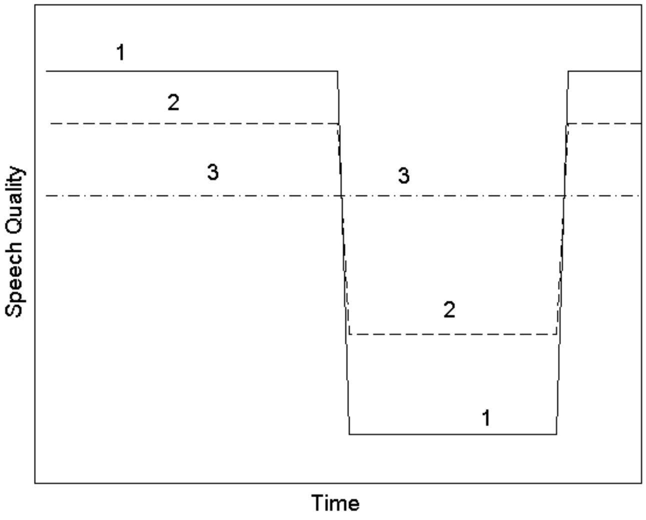

Fig. 2. Example of three levels of redundancy and three associated speech quality vs. time profiles for two-channel MDC.

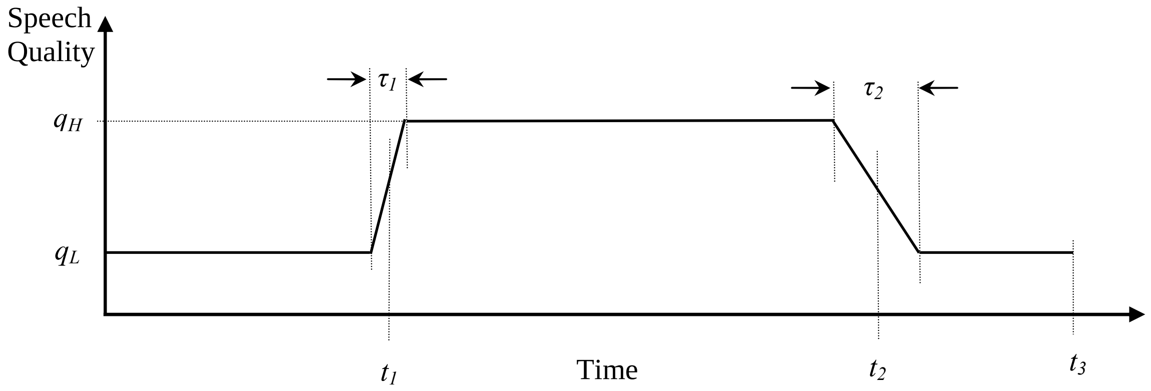

Fig. 3. A basic component of time-varying speech quality.

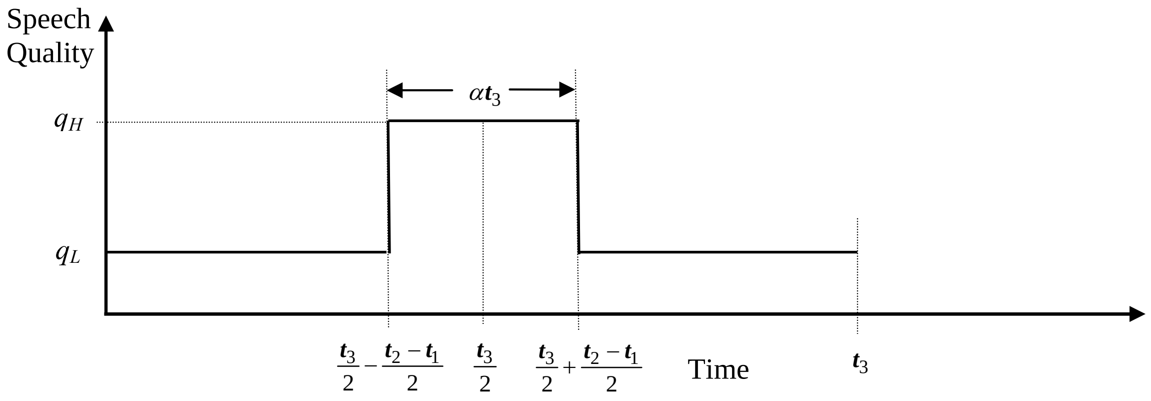
Speech Quality
$\alpha \boldsymbol{t}_3$
q_H
q_L
$\frac{\boldsymbol{t}_3}{2} - \frac{\boldsymbol{t}_2 - \boldsymbol{t}_1}{2}$
$\frac{\boldsymbol{t}_3}{2}$
$\frac{\boldsymbol{t}_3}{2} + \frac{\boldsymbol{t}_2 - \boldsymbol{t}_1}{2}$
Time
$\boldsymbol{t}_3$

Fig. 4. A simplified and constrained version of Figure 3, used in experiment conditions 1-35.

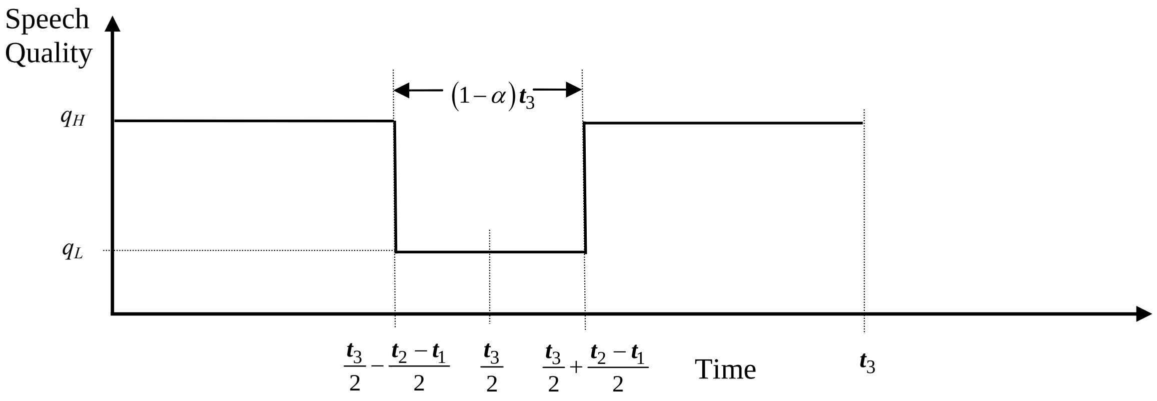
Speech Quality
$(1-\alpha)\boldsymbol{t}_3$
q_H
q_L
$\frac{\boldsymbol{t}_3}{2} - \frac{\boldsymbol{t}_2 - \boldsymbol{t}_1}{2}$
$\frac{\boldsymbol{t}_3}{2}$
$\frac{\boldsymbol{t}_3}{2} + \frac{\boldsymbol{t}_2 - \boldsymbol{t}_1}{2}$
Time
$\boldsymbol{t}_3$

Fig. 5. The speech quality history used in experiment conditions 36-40.

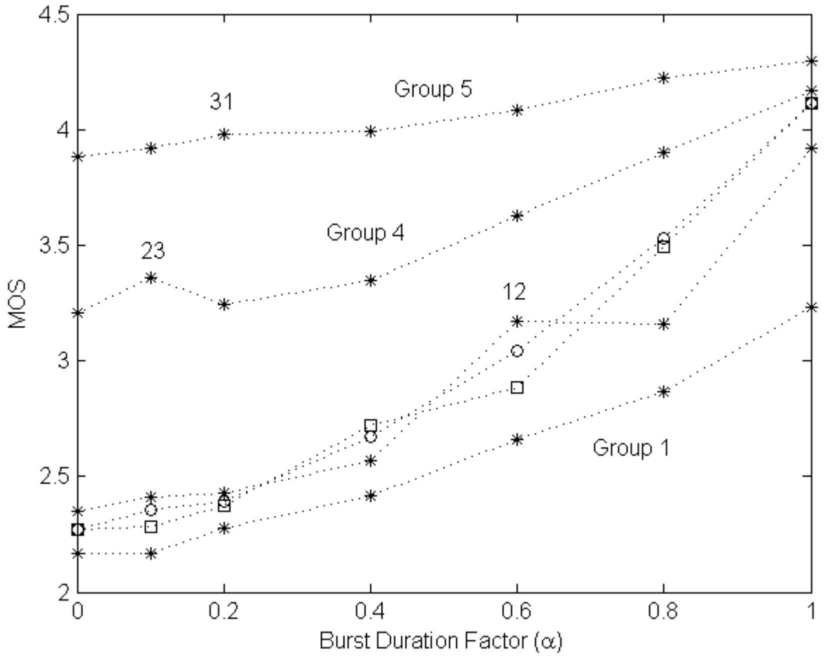

Fig. 6. MOS values vs. burst duration factor for 40 conditions organized into 6 groups. Groups 1, 4, and 5 are labeled. For unlabeled groups, asterisks indicate group 2, squares indicate group 3, and circles indicate group $\tilde{3}$. Conditions 12, 23, and 31 are labeled and are discussed in 4.4.

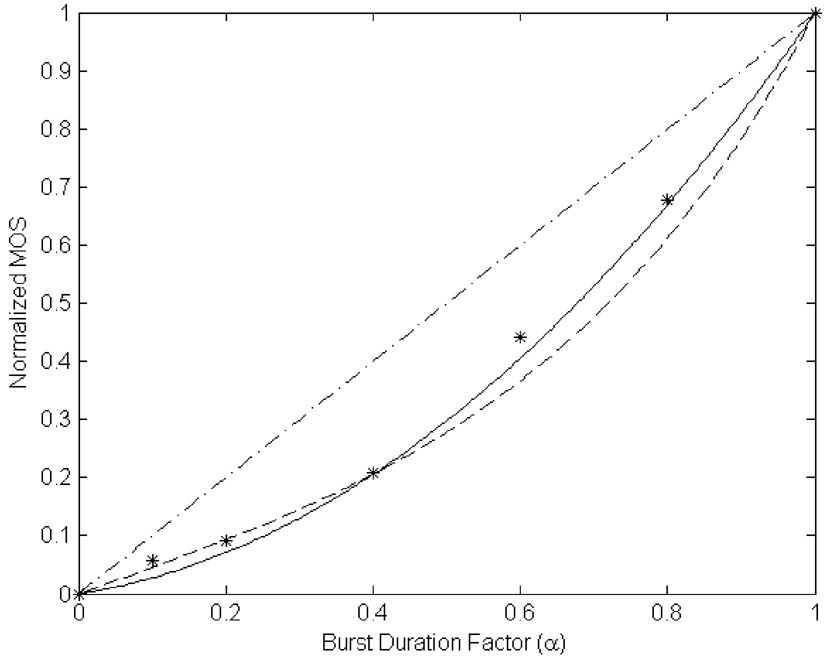

Fig. 7. Averaged MOS value, two fits, and temporal averaging in normalized space. Asterisks indicate averaged MOS values, solid line represents parabolic fit, dashed line represents hyperbolic sine fit, and dash-dot line represents temporal averaging.

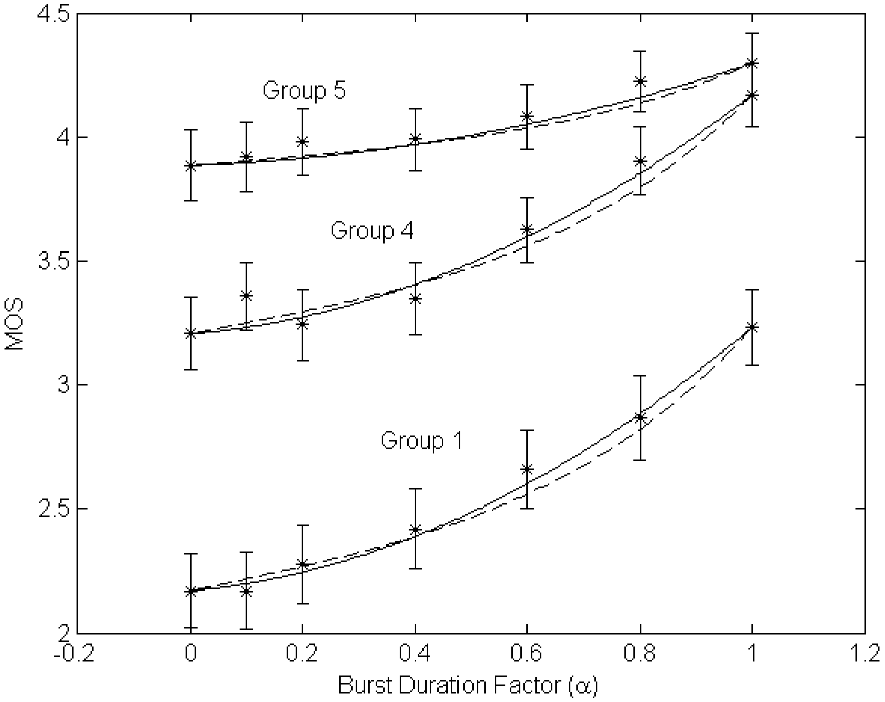

Fig. 8. MOS values, 95% confidence intervals, and two fits for groups 1, 4, and 5. Solid line represents parabolic fit and dashed line represents hyperbolic sine fit.

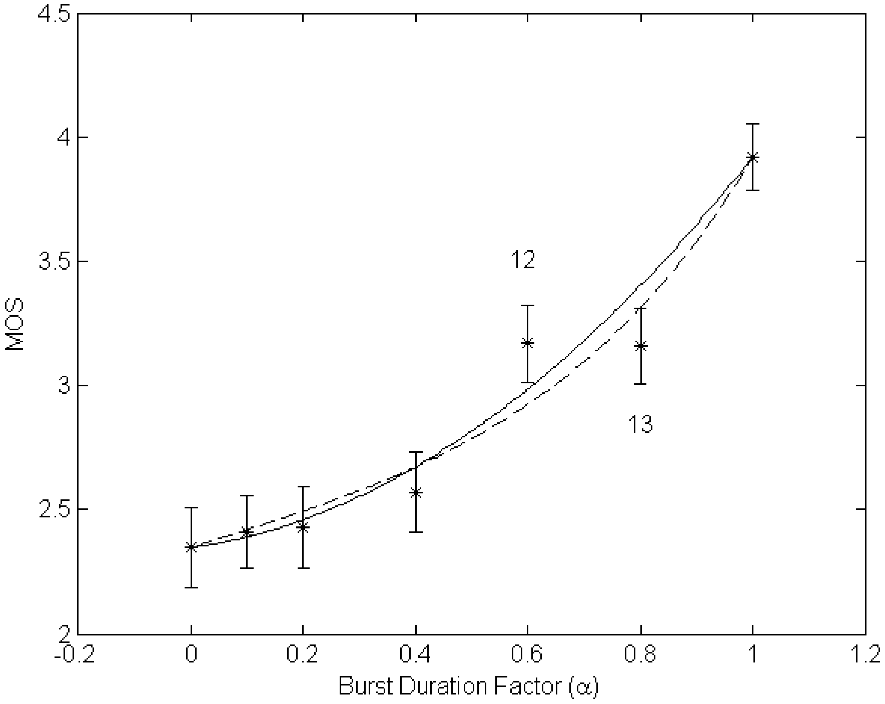

Fig. 9. MOS values, 95% confidence intervals, and two fits for group 2. Solid line represents parabolic fit and dashed line represents hyperbolic sine fit.

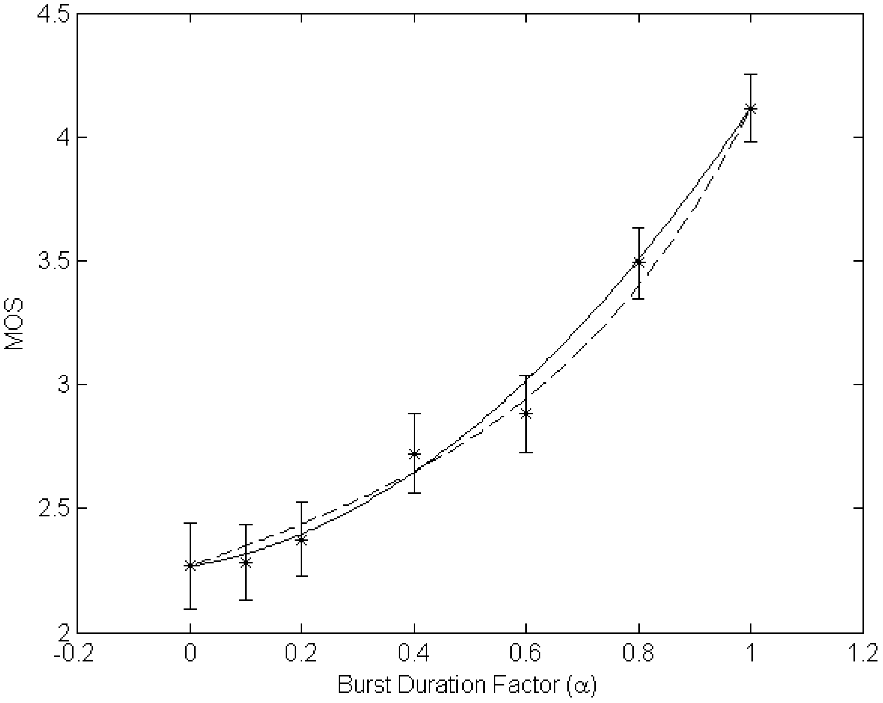

Fig. 10. MOS values, 95% confidence intervals, and two fits for group 3. Solid line represents parabolic fit and dashed line represents hyperbolic sine fit.

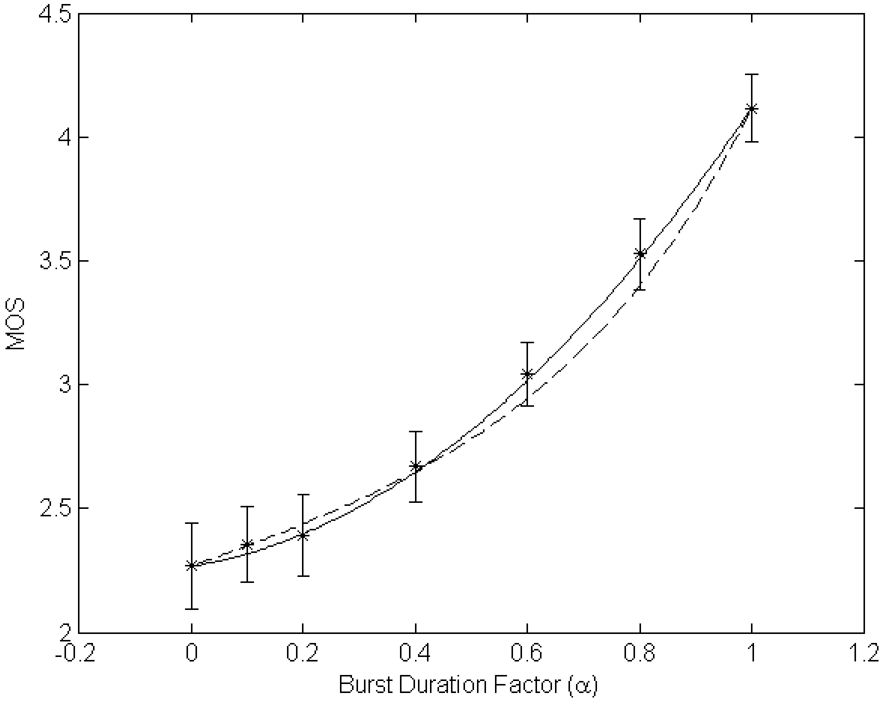

Fig. 11. MOS values, 95% confidence intervals, and two fits for group $\tilde{3}$. *Solid line represents parabolic fit and dashed line represents hyperbolic sine fit.*

Table 1. Condition and group definitions, MOS values, and 95% confidence intervals (based on 112 votes per condition).

Condition Number	Group Number	$\boldsymbol{q}_L$	$\boldsymbol{q}_H$	α	Mean	95% Confidence Interval
1	1	$\boldsymbol{q}_1$	$\boldsymbol{q}_2$	0.0	2.17	± 0.15
2	1	$\boldsymbol{q}_1$	$\boldsymbol{q}_2$	0.1	2.17	±0.16
3	1	$\boldsymbol{q}_1$	$\boldsymbol{q}_2$	0.2	2.28	±0.16
4	1	$\boldsymbol{q}_1$	$\boldsymbol{q}_2$	0.4	2.42	±0.16
5	1	$\boldsymbol{q}_1$	$\boldsymbol{q}_2$	0.6	2.66	±0.16
6	1	$\boldsymbol{q}_1$	$\boldsymbol{q}_2$	0.8	2.87	±0.17
7	1	$\boldsymbol{q}_1$	$\boldsymbol{q}_2$	1.0	3.23	±0.15
8	2	$\boldsymbol{q}_1$	$\boldsymbol{q}_3$	0.0	2.35	±0.16
9	2	$\boldsymbol{q}_1$	$\boldsymbol{q}_3$	0.1	2.41	±0.14
10	2	$\boldsymbol{q}_1$	$\boldsymbol{q}_3$	0.2	2.43	±0.16
11	2	$\boldsymbol{q}_1$	$\boldsymbol{q}_3$	0.4	2.57	±0.16
12	2	$\boldsymbol{q}_1$	$\boldsymbol{q}_3$	0.6	3.17	±0.16
13	2	$\boldsymbol{q}_1$	$\boldsymbol{q}_3$	0.8	3.16	±0.15
14	2	$\boldsymbol{q}_1$	$\boldsymbol{q}_3$	1.0	3.92	±0.13
15	$3, \tilde{3}$	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.0	2.27	±0.17
16	3	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.1	2.29	±0.15
17	3	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.2	2.38	±0.15
18	3	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.4	2.72	±0.16
19	3	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.6	2.88	±0.15
20	3	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.8	3.49	±0.14
21	$3, \tilde{3}$	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	1.0	4.12	±0.14
22	4	$\boldsymbol{q}_2$	$\boldsymbol{q}_4$	0.0	3.21	±0.15
23	4	$\boldsymbol{q}_2$	$\boldsymbol{q}_4$	0.1	3.36	±0.14
24	4	$\boldsymbol{q}_2$	$\boldsymbol{q}_4$	0.2	3.24	±0.14
25	4	$\boldsymbol{q}_2$	$\boldsymbol{q}_4$	0.4	3.35	±0.14
26	4	$\boldsymbol{q}_2$	$\boldsymbol{q}_4$	0.6	3.63	±0.13
27	4	$\boldsymbol{q}_2$	$\boldsymbol{q}_4$	0.8	3.90	±0.14
28	4	$\boldsymbol{q}_2$	$\boldsymbol{q}_4$	1.0	4.17	±0.13
29	5	$\boldsymbol{q}_3$	$\boldsymbol{q}_4$	0.0	3.88	±0.14
30	5	$\boldsymbol{q}_3$	$\boldsymbol{q}_4$	0.1	3.92	±0.14
31	5	$\boldsymbol{q}_3$	$\boldsymbol{q}_4$	0.2	3.98	±0.13
32	5	$\boldsymbol{q}_3$	$\boldsymbol{q}_4$	0.4	3.99	±0.13
33	5	$\boldsymbol{q}_3$	$\boldsymbol{q}_4$	0.6	4.08	±0.13
34	5	$\boldsymbol{q}_3$	$\boldsymbol{q}_4$	0.8	4.22	±0.12
35	5	$\boldsymbol{q}_3$	$\boldsymbol{q}_4$	1.0	4.29	±0.12
36	$\tilde{3}$	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.1	2.36	±0.15
37	$\tilde{3}$	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.2	2.39	±0.16
38	$\tilde{3}$	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.4	2.36	±0.15
39	$\tilde{3}$	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.6	3.04	±0.13
40	$\tilde{3}$	$\boldsymbol{q}_1$	$\boldsymbol{q}_4$	0.8	3.53	±0.14

Table 2. Grand means and 95% confidence intervals for conditions with nominally constant speech quality.

Speech Coder	Defined Quality Level	Conditions	Number of Votes	Mean	95% Confidence Interval
MELP, 1.2 kbps	$\boldsymbol{q}_1$	1, 8, 15	336	2.26	0.09
AMR-NB, 4.75 kbps	$\boldsymbol{q}_2$	7, 22	224	3.22	0.11
AMR-NB, 7.95 kbps	$\boldsymbol{q}_3$	14, 29	224	3.90	0.10
G.711 PCM, 64 kbps	$\boldsymbol{q}_4$	21, 28, 35	336	4.19	0.07