SOME SOURCES OF VARIABILITY IN SPEECH INTELLIGIBILITY

Ann R. Bradlow, Gina M. Torretta and David B. Pisoni
Speech Research Laboratory, Indiana University, Bloomington, Indiana, U.S.A.

ABSTRACT

Talker-specific correlates of intelligibility were explored using a large, multi-talker speech database. This database includes both sentence productions from multiple talkers and intelligibility data from multiple listeners. We examined global, talker-specific characteristics (e.g. gender, fundamental frequency, and overall speaking rate), as well as individual differences in phonetic implementation (such as vowel space compactness and fine-grained, segmental differences) as possible correlates of variation in overall talker intelligibility. Results indicated that individual differences in segmental articulation, rather than global characteristics, correlated well with overall intelligibility.

INTRODUCTION

The speech signal simultaneously encodes both linguistic and paralinguistic information [1]. Thus, in response to an utterance, a listener is made aware of both its content (the linguistic message) and of a host of information specific to the instance of the utterance. For example, due to both inter- and intra-talker differences, the speech signal conveys information about the talker's sex, age, geographical origin, physical and mental state, as well as the linguistic message he or she is trying to communicate. As a consequence of the simultaneous encoding of linguistic and paralinguistic information, we might expect an interaction between these two aspects of the acoustic signal. This study addressed this issue by investigating the correlation between talker-specific characteristics and speech intelligibility.

METHODS AND MATERIALS

The materials for this study came from the Indiana Multi-Talker Sentence Database. This database consists of 100 Harvard sentences [2] produced by 20 talkers (10 males and 10 females) of General American English. The sentences are all mono-clausal and contain 5 key words. Examples of the sentences are given below in Table 1.

Table 1. Two sample Harvard sentences with keywords underlined.

Rice is often served in round bowls.
Two blue fish swam in the tank.

Along with the audio recordings, this database includes intelligibility data in the form of sentence transcriptions by 10 listeners per talker. In the collection of these transcriptions, the listeners heard the full set of 100 sentences produced by a single talker. The listeners heard each sentence over headphones, and then typed what they heard at a computer keyboard. The sentences were presented in the clear (no noise was added) at a comfortable listening level. The listeners were all students at Indiana University with no speech or hearing impairments.

The sentence transcriptions were scored by a criterion that counted a sentence as correctly transcribed if, and only if, all 5 keywords were correctly transcribed. Any error on a keyword resulted in the sentence being counted as mistranscribed. With this scoring method, each sentence for each talker received an intelligibility score out of a possible 10. Each talker's overall intelligibility score was then calculated as the average intelligibility score across all 100 sentences. Across all 20 talkers, there was considerable variation in overall intelligibility. The intelligibility scores ranged from 81% to 93%, with a mean and standard deviation of 88% and 3%, respectively. Thus, the materials in this database covered a range of talker intelligibility and could be used as the basis for an investigation of the effect of talker-specific characteristics on overall intelligibility.

Our general approach to this investigation was to focus on two aspects of talker-specific variation. First, we examined the correlation of global talker characteristics, such as gender, overall speaking rate, and fundamental frequency, with overall talker intelligibility. Second, we looked at several aspects of the acoustic signal that provide information about the pronunciation characteristics of the talker. Specifically, we compared vowel space compactness across talkers, and performed an analysis of specific listener errors and their correlation with fine-grained talker variation at the acoustic-phonetic level.

GLOBAL CHARACTERISTICS

One of the most salient paralinguistic factors that is conveyed by the speech signal is the sex of the talker. Due to physical differences between the typical male and female vocal apparatus, as well as due to socio-linguistically determined differences between male and female pronunciation patterns, the sex of the talker is a very prominent paralinguistic factor. Furthermore, there is evidence in the literature that females tend to exhibit fewer instances of reduced speech than males [3]. Thus, we might expect female talkers to have higher overall intelligibility scores than male talkers.

In the Indiana Multi-Talker Sentence Database, the overall intelligibility scores indicated a significant sex-based difference in sentence intelligibility. The female talkers had a significantly higher average intelligibility score than the male talkers (89.4% versus 86.3%, with standard errors of 0.67% and 1.00%, respectively, $t(18)=2.57$, $p=.02$ by an unpaired, 2-tail t-test). Furthermore, in this database, the four talkers with the highest intelligibility scores were female and the four talkers with the lowest intelligibility scores were male. Thus, these data suggest that overall speech intelligibility is affected by the talker's sex. We now turn to an investigation of other paralinguistic factors that might help to explain the acoustic-phonetic reasons for this sex-based difference in intelligibility.

Overall rate of speech is a global characteristic of speech production that not only varies across talkers, but also has an impact on speech perception [4]. Using mean sentence duration as a measure of overall speaking rate, we investigated the correlation between overall rate and intelligibility. We hypothesized that slower speaking rates would correlate with higher overall intelligibility scores. However, across all 20 talkers, there was no correlation between overall rate and intelligibility. We also hypothesized that less variance in speaking rate would correlate with better intelligibility. When all talkers were included in this analysis, we found no correlation between rate standard deviation and mean intelligibility. However, when the three talkers with the lowest mean intelligibility score were excluded from the analysis, we found a high negative correlation ($R^2=-.82$) between rate standard deviation and mean intelligibility. Thus, for a subset of talkers, although mean speaking rate does not predict intelligibility, the less the variability in speaking rate the higher the intelligibility. With respect to rate differences for the males and females, we did not find that the females had generally slower rates than the males. This suggests that the sex-based difference in overall intelligibility does not result from a difference in overall speaking rate.

Another global talker characteristic that we investigated as a possible correlate of overall intelligibility was fundamental frequency. Here we hypothesized that both the mean and range of a talker's fundamental frequency might affect his or her overall intelligibility. For the male talkers, we found no correlation between mean fundamental frequency and mean intelligibility score across all 100 sentences. However, for the females, we found a moderate, negative correlation ($R^2=-.32$). Thus, these data provide some suggestion that females with lower mean fundamental frequencies might be more intelligible. With regard to fundamental frequency range, we found a moderate positive correlation ($R^2=0.38$) between F0 range and overall intelligibility for all 20 talkers, indicating that a wider range of pitch variation can enhance sentence intelligibility.

From these investigations of global talker characteristics and overall talker intelligibility, we concluded that the correlations are generally weak to moderate for the twenty, normal talkers in our database. Even though we did find a significantly higher mean intelligibility score for the females than for the males in our database, we were

unable to reliably trace this difference to global talker characteristics, such as overall speaking rate and fundamental frequency characteristics. In light of this result, we turned our attention to some of the indicators of talker variability in pronunciation. Our expectation here was that inter-talker differences at the fine-grained acoustic-phonetic level would correlate with variance in overall intelligibility.

PHONETIC IMPLEMENTATION

We began with an investigation of vowel-space characteristics. Talkers differ in the extent to which they differentiate the vowel categories in the F1 by F2 vowel space [5]. Thus, the compactness of a talker's vowel space is an indicator of the talker's pronunciation characteristics. Since a relatively expanded vowel space indicates less reduced vowels, we hypothesized that a more expanded vowel space would correlate with higher overall intelligibility.

In order to compare vowel spaces across talkers, we selected vowels from the sentence materials that provided an indication of the extremes of each talker's general vowel space. We selected three tokens of each of three point vowels, /i, u, a/. Each token came from a separate sentence, giving us a subset of nine sentences. First and second formant frequencies were measured from the steady-state portion of each of the target vowels for each of the 20 talkers. These measurements were then transformed according to the perceptually motivated mel scale, and plotted in the F1 by F2 mel space. Euclidian areas were then calculated for the triangles formed by the most extreme vowel tokens of each talker's measured vowel space.

Since these vowel space areas are representative of a subset of the total set of 100 sentences, we used the intelligibility scores across this subset of sentences in our analysis of the correlation between vowel space and intelligibility. A rank order correlation between talker vowel space area and overall intelligibility was moderately positive (Spearman rho =+0.36), indicating that across all talkers a more expanded vowel space can lead to higher overall intelligibility. Furthermore, in a comparison of the vowel space area of the male and female talkers, we found that the area within the female vowel spaces tended to be larger than the male vowel space areas (p=.037 by a 1-tail, unpaired t-test). Thus, the results of this analysis of vowel space expansion and overall intelligibility indicate that talkers who have more differentiated vowel articulations tend to be more intelligible. Furthermore, the vowel-space data suggest that the sex-based intelligibility difference might be related to sex-based differences in articulatory precision.

In order to further investigate the pronunciation characteristics that might correlate with talker intelligibility, we performed analyses of the acoustic-phonetic correlates of consistent listener errors. In these analyses we focused on specific portions of sentences that resulted in consistent listener errors, and attempted to find talker pronunciation differences that were responsible for the occurrence of listener errors.

One such case occurred in the phrase "the play seems," which was often mis-transcribed by listeners as "the place seems." In order to investigate the timing characteristics that determined the syllabification of the medial /s/, we measured the durations of the target /s/ as well as of the surrounding segments for each of the 20 talkers. We then examined the correlations between these measurements and the likelihood of correct transcription by the listeners across all talkers. Results of these measurements showed a fairly strong negative correlation (R^2=-0.65) between the duration of the medial /s/ as a proportion of the duration of the preceding word /plej/, and the rate of correct transcription. In other words, the shorter the /s/ relative to the preceding word, the more likely it was to be syllabified by listeners as onset of the following word, rather than as both coda of the preceding word and onset of the following word. Thus, in order to be correctly transcribed, this phrase required a high degree of inter-segment timing accuracy. Furthermore, there were fewer listener errors for the female productions of this phrase, indicating that the female talkers in our database were more accurate in this regard than the males.

Another case of a consistent listener error across all talkers was in the phrase, "the walled town" which was often transcribed as "the wall town." In order to explore the acoustic-phonetic factors that determined whether the word final /d/ was detected, we performed duration measurements on various portions of the target word sequence, "walled town." Results showed a positive rank-order correlation between the absolute vowel-to-vowel duration (i.e., the duration from the offset of the /a/ in "walled" to the onset of the /a^{U}/ in "town") with the likelihood of /d/ detection across all 20 talkers (Spearman rho =+0.702). However, we found an even higher correlation between rate of /d/ detection and the absolute duration of voicing during the /d/ closure (Spearman rho =+0.744). In addition to investigating the correlations of these durations in an absolute sense, we also investigated the correlation between rate of /d/ detection and these durations relative to the surrounding segment and word durations. However, the highest correlation was between absolute duration of voicing during closure and rate of /d/ detection. Since voiced stops in this pre-stop environment are typically not released, the only cue to the presence of a voiced stop is voicing during the closure. And, as demonstrated by the consistent listener error in this example, this normally variable cue can be crucial in this environment. This case is another example of talker-variation at a fine-grained, acoustic-phonetic level that has a direct effect on sentence intelligibility.

In addition to these examples of common listener errors that occurred across all 20 talkers, there were several cases of common listener errors for certain individual talker's productions of particular sentence portions. For these sentences, we compared the acoustic characteristics of the target sentence portion from the talker who was often misheard with those of a talker who received no listener errors on that sentence portion. One such instance occurred for the target phrase " smooth planks," which for one talker was often transcribed as "smooth banks." As compared to a talker whose utterance produced no listener errors on this word, this talker had a reduced /p/ closure duration, as well as a reduced /p/ VOT duration. Thus, for this talker, the cues to the unvoiced consonant were reduced in duration, resulting in listeners perceiving a voiced initial consonant.

In general, our investigations of the acoustic-phonetic correlates of specific listener errors show that variation in talker intelligibility can depend on fine-grained variation in articulation. Sentences spoken by talkers who are more precise in their articulations are more likely to be correctly transcribed.

CONCLUSIONS

The results of this investigation indicate that differences in fine-grained, articulatory-acoustic patterns correlate with variability in overall speech intelligibility. In contrast, global talker characteristics (such as mean fundamental frequency and speaking rate) are not well correlated with differences in talker intelligibility. Furthermore, this study indicated that female speakers, who tend to have more precise articulations, also have higher overall intelligibility scores than males. These findings indicate that talker-specific variations at the acoustic-phonetic level have an impact on both the paralinguistic information carried by the utterance and on its intelligibility.

ACKNOWLEDGMENTS

This work was supported by NIDCD Training Grant DC-00012 and by NIDCD Research Grant DC-00111 to Indiana University in Bloomington, IN.

REFERENCES

[1] Laver, J. & Trudgill, P. (1979), "Phonetic and linguistic markers in speech." In K. Scherer & H. Giles (Eds.), *Social Markers in Speech.* Cambridge, UK:Cambridge Univ. Press.

[2] IEEE (1969), "IEEE recommended practice for speech quality measurements," *IEEE Report No. 297.*

[3] Byrd, D. (1994), "Relations of sex and dialect to reduction," *Speech Communication* vol. 15, pp. 39-54.

[4] Miller, J. (1987), "Rate-dependent processing in speech perception." In A. Ellis (Ed.), *Progress in the Psychology of Language.* Hillsdale, NJ: Erlbaum.

[5] Bond, Z. and Moore, T. (1994), "A note on the acoustic-phonetic characteristics of inadvertently clear speech," *Speech Communication* vol. 14, pp. 325-337.