AUGMENTED INTELLIGIBILITY IN SIMULTANEOUS MULTI-TALKER ENVIRONMENTS

Nima Mesgarani
Shihab Shamma

Neural System Lab.
Institute for System Research
University of Maryland
mnima,sas@glue.umd.edu

Ken W. Grant

Army Audiology and
Speech Center, Walter Reed
Army Medical Center
grant@tidalwave.net

Ramani Duraiswami

Perceptual Interfaces and
Reality Lab., UMIACS
University of Maryland
ramani@umiacs.umd.edu

ABSTRACT

Speech intelligibility can be severely compromised in environments where there are several competing speakers. It may be possible, however, to improve speech intelligibility in such environments by manipulating certain acoustic parameters known to facilitate the segregation of competing signals. The first step in designing such a feature-rich audio display is to understand the significant elements of human auditory perception that affect information transmission capacity. We review a series of experiments to examine the impact of different audio-display design parameters on overall intelligibility of simultaneous speech messages and show how using these features may improve intelligibility.

1. INTRODUCTION

Speech communication seldom takes place under pristine conditions. More often, environmental noise, reverberation, or competing signals from other speakers can interfere with the accurate transmission and reception of speech. One communication situation commonly encountered in the real world, known as the "cocktail party" effect [1] has received a great deal of attention among speech scientists in recent years. For decades, researchers have questioned how it is possible for listeners to separate out different overlapping voices arriving at the two ears simultaneously and selectively attend to the desired message while tuning out others. There are some circumstances, however, where listeners must attend to more than one message at a time. Aircraft pilots, for example, typically monitor several voice channels simultaneously, receiving information from co-pilots as well as ground crews. The purpose of this research is to develop and evaluate a customizable, audio-user interface capable of displaying multiple acoustic inputs in a manner that enhances the listeners' abilities to understand each separate message.

Although little research has been done regarding the problem of speech reception of multiple, simultaneous messages, there has been substantial research dealing with the problem of selective attention of one speech message against a background of other competing messages. The original problem posed by Cherry [1] was to identify the factors used to separate what one person is saying when others are speaking. A number of possible factors were identified. These included the location of each speaker, visual speech cues (e.g. those derived by speechreading), different voice characteristics (e.g. pitch, speed, gender, and accent), and linguistic properties of each separate message (e.g. lexical and sentential constraints). In one study, listeners had to verbally repeat back the words or phrases from two mixed speeches recorded on tape and played back over a single loudspeaker. Subjects were able to play the tape as many times as required. When the subject matter of the two passages was distinctly different, listeners had little difficulty reconstructing the phases, but in some instances they had to replay the tape 10-20 times before disentangling the two streams correctly. However when the passages were constructed out of common clichés, separating the two competing messages became much more difficult. In a second study [1], the competing speech passages were fed separately to the left and right ear of a headphone. In this case, subjects had no difficulty listening to either speaker, switching at will between messages, regardless of the semantic and syntactic structure of the passage. However, while attending to one speaker, the subjects were mostly unaware of the competing message. That is, they were unable to report any words, identify the gender of the speaker, or language, or even if reversed speech was heard in the competing ear. In short, specific details of the "rejected" signal went mostly unnoticed.

In recent years, the cocktail party effect has been recast as a problem of source segregation and fusion [2]. An extensive body of psychophysical work exists regarding the ability of the human auditory system to segregate sounds streams based on temporal and spectral differences, and to group sounds based on temporal and spectral coherence (see [3] and [4] for reviews). These studies have increased our understanding of basic auditory perceptual mechanisms making it possible to enhance a listener's ability to segregate sound streams by manipulating certain acoustic parameters such as source localization [5-11], pitch [12-15], level [16,17], timbre [4,18,19], and temporal asynchrony [20]. By enhancing differences along certain dimensions of voice quality, timing and spatial location, it may be possible to overcome some of the limitations observed by Cherry [1] regarding a listener's ability to monitor and recognize more than one sound source at a time. For example, if a pitch difference is superimposed onto two sentences spoken concurrently by a single male speaker, recognition accuracy was shown to increase from roughly 60% correct (with no pitch separation) to 75% correct (with a separation of eight

semitones). The studies by Assmann [15] and Yost et al. [21] are most germane for the current paper because listeners were required to monitor all incoming messages. This is in contrast to the vast majority of studies exploring the "cocktail party" effect, which could be characterized more accurately as employing methodologies appropriate for studying selective attention (i.e., attending to only one speech target) rather than divided attention. Nevertheless, it is clear that signal manipulation, which serves to help distinguish one sound from another, also helps listeners identify sound sources and their content. Little work has been done to explore the effects of combining signal manipulations, such as using spatial separation and pitch in tandem.

2. METHODS

2.1. Subjects

Five listeners with prior experience in psychoacoustics experiments served as subjects. All subjects had self-reported normal hearing. Their ages ranged from 24 to 29.

2.2. Stimuli

The speech stimuli were taken from the TI 46 Word Speech Database, NIST Speech Disc 7-1.1, September 1991. This corpus consists of speaker dependent digits from 0 to 9, which are sampled at 12.5 kHz. For each digit, the waveform was extracted manually and adjusted to a predefined time duration of 508 ms preserving pitch to eliminate differences in duration across digits and speakers. Each digit was produced five times so the subjects would be less likely to become familiar with any idiosyncratic cues that might exist in any one utterance.

Test blocks consisted of 20 trials in which three speakers each said one digit. The digits were generated randomly with 2 constraints:

(1) All three digits were different in a given trial.

(2) Each speaker produced each digit (0-9) exactly twice per block.

The stimuli were presented over headphones at a comfortable listening level (approximately 70 dB). The listeners' task was to report *all three* of the digits in a given trial correctly for a correct response to be noted. Responses were collected using a computer mouse to select the appropriate numbers on the screen. The tests were conducted using a laptop computer in a sound treated room.

Individual head-related transfer functions (HRTFs) were used for spatial processing of sounds. The HRTFs were measured on 1037 points around the head. Both front and back hemisphere HRTFs were measured over a 10-degree grid in double polar coordinates as described in [22]. For conditions which employed spatial location as a potential cue for enhancing intelligibility, the speech signals for each speaker were filtered with the HRTFs corresponding to the desired locations and rendered through headphones.

Three primary conditions were tested.

1. Baseline tests with minimal separation between competing speakers. In these tests, all three digits were spoken by one person talking from the same azimuth and elevation (0°). Five separate speakers were tested (two male, three female).
2. Timbre and Pitch effects were tested by replacing the one speaker in the baseline condition with a mixture of female and male speakers each having a different average fundamental frequency and timbre.
3. Spatial Separation was tested in four different conditions. For the first three tests, three different dimensions (azimuth, elevation and range) were tested independently. For the forth test, azimuth and elevation dimensions were used in tandem.

The listeners first participated in a block of 20 practice trials for each test, with feedback after each trial, prior to data collection. To minimize the undesired effects of training, the three primary test conditions (baseline, timbre and pitch, and spatial separation) were presented alternately during each test session.

3. BASELINE MEASURES

Baseline tests included five different speakers (two male and three female) with average fundamental frequencies (F0) ranging between 132 to 224 Hz. Figure 1 shows that there was a decrease in intelligibility as the speaker's F0 increased. In other words, the masking effect of competing speakers seemed to be more effective as F0 increased. This phenomenon was also observed in informal tests with a single male speaker with pitch shifted by 4 and 8 semitones. We are unaware of either a similar finding in the literature or a simple explanation for this result. In order to allow a comfortable range for performance improvement, we chose the female speaker with the highest fundamental frequency as our baseline (Speaker F3). Subsequent tests examined the effects of separation in pitch, timbre and location, both separately and in combination, to determine if recognition performance for three simultaneously spoken digits could be improved over baseline.

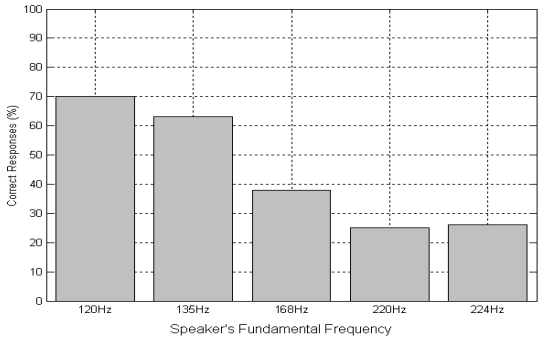

Figure1. *Percentage of correct responses for different baseline tests using speakers with different fundamental frequencies. The baseline tests consisted of three simultaneous digits all said by the same person. Five different speakers with different fundamental frequencies were used.*

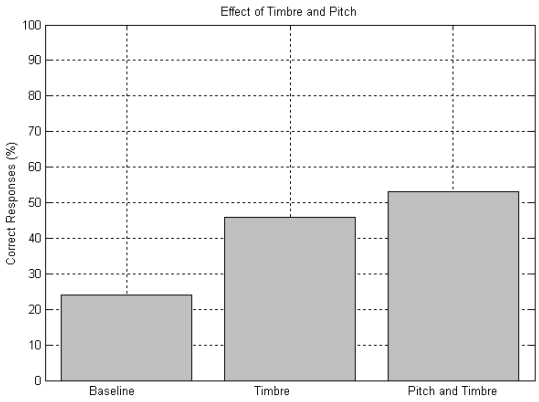

Figure 2. *Percentage of correct responses due to separation of timbre and pitch. Timbre separation was accomplished by replacing the baseline speaker (F3) with different female speakers such that their pitches were almost the same. For tests in which timbre and pitch separation were used together, speakers were selected such that they had different pitches by four and eight semitones from the original female speaker.*

4. PITCH AND TIMBRE

Differences in pitch and timbre provide an important cue that can be used to segregate competing speech signals. The voice of different talkers can vary in a wide number of ways, including differences in fundamental frequency (F0), formant frequency, and timbre. In one test condition, three *different* female speakers were used. They all had roughly the same average F0 (217, 220 and 224Hz) but different timbres. In a second test, two female and one male speakers were selected from the database, such that the pitch of the female speakers were almost four semitones apart, and the pitch of male speaker was almost eight semitones lower than the female speaker with the highest F0 (224, 168 and 135Hz). In this case, both pitch and timbre were different for competing speech signals. Figure 2 shows the effects of timbre only, as well as pitch and timbre together, on the percentage of correct responses.

5. SPATIAL SEPARATION

Spatial separation is known to be one of the most useful cues for segregating competing sound signals. By taking advantage of the subjects' HRTFs, a series of five tests were conducted in which the competing speakers were all the same person but positioned at different locations in space. In the first test, using the level difference together with HRTFs, the speakers were all located at an azimuth and elevation of 0° but at different apparent distances (i.e., range). In the second test, speakers were located at the same elevation and range, but different azimuths (-85°, 0°, 85°). In the third test, speakers were located at the same azimuth (0°) and range, but different elevations with respect to horizontal plane, (-30°, 30°, 90°). The fourth test was similar to the third, but the elevations were at 0°, 90°, and 180° (front, top and back).

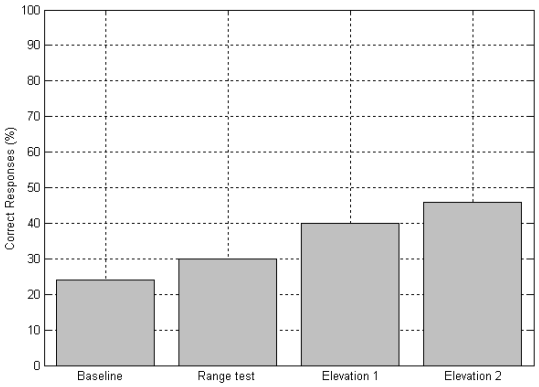

Figure 3. *Percentage of correct response adding spatial features to the speakers' voice in range and elevation dimension. In the range test, the speakers were located at different distances from the subject but at the same azimuth and elevation (0°). In Elevation 1, speakers were located at an azimuth of 0°, and at elevations of 0°, 90°, and 180°. Elevation 2 test was similar to Elevation 1 but with elevations at -30°, 30°, and 90°.*

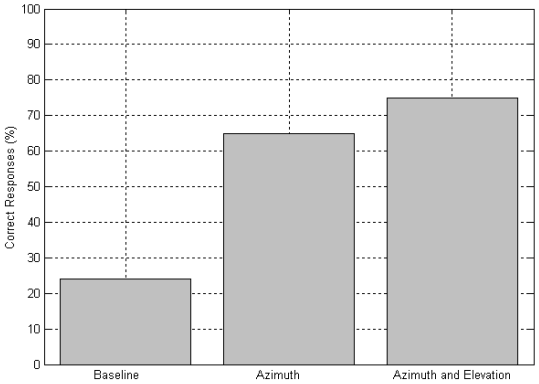

Figure 4. *Percentage of correct response adding full spatial separation to the speakers voice first in azimuth dimension only, and second, in both azimuth and elevation.*

In the last test, using both azimuth and elevation, speakers were located at (az –60°, el 0), (az 0°, el 60°), and (az 60°, el 0°). Figure 3 illustrates the effect of elevation and range separation on the percentage of correct response. A small improvement in intelligibility occurred using range as a segregation cue, whereas more substantial improvements were observed with elevation as the cue. Figure 4 shows the effect of using separations in azimuth to disentangle the competing speech signals. As expected, azimuth provided greater enhancement to intelligibility than did elevation (which is basically a special case of a timbre change), or range. With both azimuth and elevation cues available, the best result were obtained as shown in Figure 4.

6. CONCLUSIONS

An efficient way to augment the intelligibility in multi-talker environments is to use virtual synthesis techniques to separate the voice characteristic and location of the competing speakers. The data from Figure 1 demonstrate that the intelligibility of multiple competing speakers may depend on the speaker's average fundamental frequency. Speakers with higher average F0 tend to exhibit greater self-masking. Baseline intelligibility scores for male speakers with lower F0 were consistently better than those obtained for female speaker with higher-pitched voices. These results should be viewed with caution however, because of the method used to implement the change in F0 (selecting different talkers also varied other aspects of the voice). In pilot tests with male speakers (not shown here), there was only a slight improvement when timbre cues were used to segregate competing speakers. Basically, male and female multi-talker intelligibility scores were comparable when enough spatial separation was added to the baseline condition. Additional research is necessary to explore these phenomena in more detail. Spatial separation turned out to be the most effective single cue for improving the intelligibility of multiple, competing speakers. As shown in Figures 3 and 4, this improvement is not limited to the use of the azimuth dimension (which is easily applicable using ILDs and ITDs). Elevation cues also provided significant benefits to multi-talker intelligibility. However, rendering accurate elevation cues is computationally and practically more expensive than azimuth because it requires individual HRTFs.

Although this paper has reviewed many of the factors that can influence the ability of listeners to recognize multiple competing speech messages, further study is needed to determine how the different cues outlined in this paper interact with one another. Also, although the separate effects of localization, azimuth, elevation, and range were reviewed in this paper, more research is needed to study the optimal location for each speaker in space.

7. ACKNOWLEDGEMENTS

Partial support of ONR Award N000140210571 and NSF Grant 0205271 is gratefully acknowledged. We would also like to thank Dr. Elena Grassi and Dr. Dimitry Zotkin for providing the measured HRTFs.

8. REFERENCES

[1] Cherry, E.C. (1953). "Some experiments in the recognition of speech, with one and two ears," *J. Acoust. Soc. Am*. 25, 975-979.

[2] Bregman, A.S. (1991). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, Massachusetts: MIT Press.

[3] Darwin, C.J., and Carlyon, R.P. (1995). "Auditory grouping," in B.C.J. Moore (ED, *The Handbook of Perception and Cognition*. Volume 6, Hearing, Academic Press: London, pp.387-424.

[4] Ericson, M.A., and McLinley, R.L. (1991). "The intelligibility of multiple talkers separated spatially in noise," in R.H. Gilkey and T.B. Anderson (Eds), *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum: Hillsdale, NJ. p.p. 701-724.

[5] Kidd, G.J., Mason, C.R., and Rohtla, T.L. (1995). "Binaural advantages for sound pattern identification" *J. Acoust. Soc. Am*. 98, 1977-1986.

[6] Kidd, G.J., Mason, C.R., Rohtla, T.L., and Deliwala, P.S. (1998). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns" *J. Acoust. Soc. Am* 104, 422-431.

[7] Bird, J., and Darwin, C.J. (1998). "Effect of a difference in fundamental frequency in separating two sentences," In A.R. Palmer, A. Rees, A.Q. Summerfield, and R Meddis (Eds.), *Psychophysical and Physiological Advances in Hearing*. London: Whurr.

[8] Freyman, R.L., Helfer, K.S., McCall, D.D., and Clifton, R.K. (1999). "The role of perceived spatial separation in the unmasking of speech," *J. Acoust. Soc. Am*. 106, 3578-3588.

[9] Freyman, R.L., Balakrishnan, U., and Helfer, K.S. (2001). "Spatial release from informational masking in speech recognition," *J. Acoust. Soc. Am*. 109, 2112-2122.

[10] Hawley, M.L., Litovsky, R.Y., and Colburn, H.S. (1999). "Speech intelligibility and localization in a multi-source environment" *J. Acoust. Soc. Am*. 105, 3436-3448.

[11] Drullman, R., and Bronkhorst, A.W. (2000). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation," *J. Acoust. Soc. Am*. 107, 2224-2235.

[12] Brokx, J.P.L., and Nooteboom,. S.G. (1982). "Intonation and the perceptual separation of simultaneous voices," *J. Phonetics*, 10, 23-36.

[13] Assman, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *J Acoust. Soc. Am*. 88(2), 680.

[14] Summers, V., and Leek, M.R. (1998). "F0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss*," J. Sp. Lang. Hear. Res*. 41, 1294-1306.

[15] Assmann, P.F. (1999). "Fundamental frequency and the intelligibility of competing voices," Proceeding of the 14th International Conference of Phonetic Sciences, 179-182.

[16] Rose, M.M., and Moore, B.C.J. (2000). "Effect of frequency and the level on auditory stream segregation," *J. Acoust. Soc. Am*. 108, 1209-1214.

[17] Brungart, D.S. (2001). "Informational and energetic masking effects in the perception of two simultaneous talkers," *J. Acoust. Soc. Am*. 109, 1101-1109.

[18] Darwin. C.J. and Hukin, R.W. (2000). "Effects of reverberation on spatial, prosodic, and vocal-tract size cues to selective attention," *J. Acoust. Soc. Am*. 108, 335-342.

[19] Darwin, C.J. and Hukin, R.W. (2000). "Effectiveness of spatial cues, prosody and talker characteristic in selective attention," *J. Acoust. Soc. Am*. 107, 970-977.

[20] Summerfield, Q., and Assmann, P.F. (1991). "Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony," *J. Acoust. Soc. Am*. 89(3), 1364-1377.

[21] Yost, W. A., Dye, R.H., and Sheft, S. (1996). "A simulated 'Cocktail Party' with up to three sound sources," Perception and Psychophysics. 58, 1026-1036.

[22] Grassi, E., Tulsi, J. and Shamma, S. (2003). "Measurement of Head-Related Transfer Functions based on the empirical transfer function estimate" to be presented at ICAD 2003.