THE EFFECTS OF F0 MANIPULATION ON THE PERCEIVED DISTANCE OF SPEECH

Douglas S. Brungart
Alexander J. Kordik

1Air Force Research Laboratory
and 2Sytronics, Inc.
Wright-Patterson AFB, OH 45433

Koel Das
Arnab K. Shaw

Department of Electrical Engineering
Wright State University
Dayton, OH 45435

ABSTRACT

Recent research has shown that the apparent distance of speech increases systematically with the vocal effort level of the talker. At this point, however, it is not clear how the different acoustic features of speech that change with vocal effort contribute to this increase in apparent distance. In this experiment, pitch-synchronous overlap and add (PSOLA) techniques were used to modify the F0 contours of pre-recorded speech signals to match the F0 contours of speech samples that were produced at 6 dB higher output levels and at 6 dB lower output levels. The other characteristics of the speech signals were left unchanged. Psychoacoustic tests of the resulting speech signals show that the synthetic F0 shifts could account for some, but not all, of the apparent changes in distance that occur with a 6 dB change in the output level of the talker.

1. INTRODUCTION

Although human listeners generally have difficulty judging the distances of unfamiliar sound sources, they are able to judge the distances of speech sounds relatively accurately [1, 2]. These judgments are possible because the acoustic characteristics of speech change systematically with the vocal effort level of the talker. From past experience, listeners know that whispered speech is produced at low output levels, that conversational speech is produced at intermediate output levels, and that shouted speech is produced at high output levels. By comparing the apparent production level of a speech signal with the intensity of the sound reaching the ears, it is possible to make reasonably accurate judgments about the distances of live talkers. It is, for example, relatively easy to determine that a loud-sounding whisper was produced by a closer talker than a quiet-sounding shout. Apparent distance judgments for speech sounds may also be influenced by a priori knowledge about the vocal effort adjustments that talkers make to accommodate conversations at different distances [3]. Nearby talkers tend to produce whispered speech, and far-away talkers tend to produce shouted speech. Thus, it is not surprising that listeners tend to associate whispered speech with closer talker locations than shouted speech.

Recent experiments have shown that apparent distance increases much faster with the output level of the talker for shouted speech (where it doubles with each 8 dB increase in production level) and conversational speech (where it doubles with each 15 dB increase in production level) than for whispered speech (where it is roughly independent of production level) [2]. This suggests that listeners judge the apparent distances of speech sounds on the basis of acoustic features that only occur in voiced speech or are more sensitive to production level in voiced speech than in whispered speech. One production-level dependent feature that is unique to voiced speech is the systematic increase in fundamental frequency (F0) that occurs when talkers increase the intensity of their voices. This increase in F0 is a direct result of the increase in pressure applied to the vocal folds by shouting talkers [4]. Although there is reason to believe that this systematic increase in F0 has some influence on the apparent distance of speech, the extent of this influence is not known. In this paper, we describe an experiment that used PSOLA techniques to modify the F0 characteristics of pre-recorded speech samples to match the F0 characteristics of speech samples produced at different output levels by the same talker. Listeners were then asked to judge the apparent distances of the original and F0-shifted speech samples. This procedure allows a direct examination of the role that F0 plays in determining the apparent distance of speech.

2. METHODS

2.1. Collection of speech stimuli

The speech samples used in the experiment were collected with a B&K 4144 microphone that was mounted one meter in front of the talker’s mouth in a large anechoic chamber.
The utterances were digitally recorded with a Tucker-Davis DD1 D/A, and displayed on the screen of the control computer. Then an experimenter visually inspected the recording for any clipping and truncated the recording to the beginning and end of the utterance. The RMS power of the truncated signal was then calculated and compared to the RMS power of a 94 dB, 1 kHz calibration tone to determine the production level of the utterance. If the utterance was produced within 1 dB of the desired production level, it was saved and the talker was instructed to produce the next utterance. If the utterance was not produced within 1 dB of the desired production level, the talker was instructed to increase or decrease his or her production level and repeat the utterance. Four talkers (two male and two female) were used to collect four different utterances selected from a 30-word phonetically balanced word list (“thug,” “staff,” “snipe,” and “goose”) at seven different production levels (60, 66, 72, 78, 84, 90 and 96 dB SPL). Thus a total of 112 different utterances were recorded for the experiment.

2.2. F0 Manipulation

These 112 utterances were used to produce 196 additional F0-shifted speech stimuli. The utterances produced at levels from 60 dB to 90 dB were used to produce “up-shifted” stimuli with F0 contours matching the contours of the same utterances produced at a 6 dB higher production level. The utterances produced at levels from 66 to 96 dB were used to produce “down-shifted” stimuli with F0 contours matching the contours of the same utterances produced at a 6 dB lower production level.

The speech enhancement was carried out with the PRAAT software package [5]. First, the speech file having the longest duration was selected as the reference. Then the other speech files were modified in length to make their duration equal to that of the reference. Then the source speech signal (the original speech file) and the target speech signal (the +6 dB or -6 dB signal with the desired F0 contour) were manually adjusted to temporally align their voiced regions.

Next, the F0 contour of the target (+6 dB) speech signal was analyzed using the autocorrelation method contained in the PRAAT speech toolkit. The PRAAT toolkit uses the F0 contour determined from this autocorrelation method (frequency and voiced/unvoiced decisions) to position glottal pulses according to the distribution of acoustic energy in the original target signal.

Once the F0 contour of the target signal was determined, a pitch-synchronous overlap and add (PSOLA) algorithm was used to synthesize a new speech signal with the F0 contour calculated for the target (+6 dB) speech signal and the other properties of the source speech file. First, the F0 contour calculated for the target speech signal was used to generate an excitation pattern of glottal pulses along the entire length of the source file. Those pulses that fell within the unvoiced regions of the source file were eliminated and replaced with a direct copy from the original source file. For each point in the voiced interval, the nearest glottal pulse location in the original sound was found, and a piece of the original sound centered around that pulse was copied to the synthesized sound at that point using a Gaussian window. This manipulation resulted in a signal with the same F0 contour as the target speech signal and the other properties of the original source file. Figure 1 shows the average F0 values of the original, up-shifted, and down-shifted versions of the utterance “goose” for one of the two female talkers used in the experiment.

2.3. Psychoacoustic Evaluation

Six normal-hearing paid volunteers (two female, four male) with prior experience in acoustical research participated in the study. The participants sat outdoors in an open field with an array of parachute-material covered metal mesh cones directly in front of them at distances of 0.5, 1, 2, 4, 8, and 16 m. Each cone was marked with large number ranging from 1 (for the 0.5 m location) to 6 (for the 16 m location). Two of the cones (located at 2 m and 4 m) concealed battery-powered wireless Advent AW820 loudspeakers. The other four cones concealed dummy loudspeakers. Figure 2 shows the experimental setup used in the psychoacoustic experiment.

The experiment was controlled by a notebook computer that was held in the participant’s lap. The two audio output channels of the computer were connected to two RF modulators that transmitted signals to the two wireless loudspeakers used in the experiment. Prior to the experiment, the two loudspeakers were equalized to produce a 1 kHz tone at the same level (66 dB SPL) at the location of the listener (as measured with a Rion noise meter). Once this calibra-
tion was completed, the participants were brought outside and instructed that loudspeakers were located under the six numbered cones. They were then asked to listen to a series of speech stimuli and respond by pressing the number key (1-6) of the cone corresponding to the apparent location of each stimulus. They were also instructed to use half steps (e.g., 5.5) to respond to sounds that appeared to originate between two marked locations. Note that all of the speech files were equalized to the same RMS level as the 1 kHz calibration tone prior to presentation. Thus, all the sound files were presented at an overall level of approximately 66 dB SPL at the location of the listener and no intensity cues were available for the distance identification task.

Each participant responded to a total of 608 trials in a full factorial design with 4 utterances (“thug,” “staff,” “snipe,” and “goose”), four talkers, seven production levels, three F0 conditions (unshifted, up-shifted, and down-shifted), and two originating loudspeaker locations (2 m or 4 m). These trials were collected in random order in 12 blocks of 50 trials, plus an additional 13th block of 8 trials plus any repeats of trials where the listener was unable to respond due to environmental noise.

3. RESULTS

The overall results of the experiment are summarized in Figure 3. These results clearly illustrate the strong relationship that exists between the production level of a speech signal and its apparent distance. The results for the unshifted condition (open circles) are consistent with previous results [2] that have shown that the apparent distance of a shouted speech signal (produced above 72 dB) doubles with each 8 dB increase in production level, and that the apparent distance of conversational speech (produced below 72 dB) is much less sensitive to the production level of the talker than shouted speech.

The results from the F0-shifted conditions (dotted and dashed lines) confirm that the apparent distances of the speech signals were altered significantly by the F0 shifts introduced in the experiment. In general, down-shifting had more of an impact on apparent location than up-shifting. The down-shifting operation can account for nearly all of the reduction in apparent distance caused by a 6 dB decrease in production level at 96 dB and 90 dB, and about half of the reduction caused by a 6 dB decrease in production level at 84 dB and 78 dB. When the speech was produced at a conversational level (72 dB or 66 dB), down-shifting had no effect on apparent distance.

The up-shifting operation produced a consistent increase in the apparent distances of the speech samples across all the different production levels tested. However, this increase could only account for a small portion (less than half) of the increase in apparent distance caused by a true 6 dB increase in production level.

One of the more striking features of the results is the extent to which the production level cues dominated the natural auditory distance cues associated with the two actual loudspeaker locations tested in the experiment. Figure 4 shows the overall results of the experiment for the two actual loudspeaker locations. In each case, the mean response distances increased systematically from less than 1 m to more than 8 m as the production level of the speech increased from 60 dB to 96 dB. This was true despite the fact that all of the stimuli were actually presented from a loudspeaker at a single fixed distance in the middle of this range. Note that actual loudspeaker location had an impact.
on the apparent locations of the speech signals only when the production level was less than 72 dB; when the production level was greater than 72 dB, actual loudspeaker location had no impact on performance. A likely explanation for this result is that the relatively weak speech-based distance cues for conversational-level speech encouraged the listeners to rely on non-speech auditory distance cues in their distance judgments, while the relatively strong speech-based distance cues in the shouted speech made it unnecessary to pay attention to the non-speech distance cues in the stimulus.

Before concluding, it should be noted that the results shown in Figures 3 and 4 were very consistent across the different listeners, talkers, and utterances used in the experiment. There was almost no difference between the distance judgments for the four different utterances used in the experiment, and the mean distance judgments varied by only about 30% across the four talkers used in the experiment. These results are consistent with previous results that have shown that utterance and talker account for a much smaller proportion of overall variance than production level in judgments about the distances of speech signals [2]. The variations across the different listeners were somewhat larger, but all of their distance judgments exhibited the systematic increase with production level seen in the overall results in Figures 3 and 4.

4. CONCLUSIONS

Two important conclusions can be drawn from the results of this experiment. The first major conclusion is that F0 plays an important role in determining the apparent distances of speech sounds, and that F0 can account for some, but not all, of the variations in apparent distance that occur when the vocal effort of the talker changes. This suggests that the apparent distance of speech is also influenced by non-pitch acoustic features such as spectral tilt that also change systematically with the vocal effort level of the talker [4]. The second major conclusion is that vocal-effort-based distance cues are capable of producing compelling distance illusions that dominate the acoustic distance cues that are naturally available for sound sources located at different distances in the median plane. Production level had a much greater impact on the distance judgments than the actual presentation distances of the speech stimuli, and variations in production level resulted in distance judgments that were both closer and farther away than the actual loudspeakers used to present the stimuli in the experiment.

Together, these two conclusions have important implications for the use of speech-based distance cues in virtual audio displays. While F0 manipulation is clearly not sufficient to control the apparent distance of speech, there is reason to believe that more advanced speech processing techniques may someday be able to modify arbitrary speech signals such that they appear to originate from any desired distance relative to the listener. This would allow virtual display designers to present real-time speech signals (signals from a radio, for example) at any desired location in a virtual display. Such a capability would greatly enhance the effectiveness of virtual audio displays, which are currently very limited in their ability to provide robust, intuitive distance information.

5. REFERENCES