The Influence of Vocal Effort on Human Speaker Identification

Douglas S. Brungart†, Kimberly R. Scott†, and Brian D. Simpson††

† Air Force Research Laboratory and ‡ Veridian
Wright-Patterson AFB, OH
douglas.brungart@wpafb.af.mil

Abstract

Although many of the acoustic cues used for speaker identification change systematically with the voice level of the talker, little is known about the influence that vocal effort has on the identification of individual talkers by human listeners. In this experiment, listeners were trained to identify four different same-sex talkers speaking at one of three different levels of vocal effort (whispered, conversational, or shouted). They were then tested on their ability to identify the same four talkers speaking at the other two levels of vocal effort. The results show that the whispering talkers were substantially harder to identify than the conversational talkers, and that the conversational talkers were substantially harder to identify than the shouting talkers. The results also show that listeners who were trained to identify individual talkers speaking at one level of vocal effort had difficulty identifying the same talkers when they were speaking at a different level of vocal effort. These results confirm that changes in vocal effort have a dramatic impact on human speaker identification, and suggest that the effects of vocal effort should be considered in the design of automatic speaker identification systems.

1. Introduction

The human speech signal is characterized by a great deal of variability along a number of acoustical parameters, which co-vary with factors such as individual talker characteristics, speaking rate, phonation, and vocal effort. Despite this variability, human listeners possess a remarkable ability to identify familiar voices in a variety of contextual settings that far exceeds the ability of complex speech recognition algorithms. Determining the parameters mediating this ability has been the focus of much research. Most studies approach this problem by utilizing resynthesis techniques to generate speech signals that have been modified along one, or multiple, acoustic dimensions. The results from these studies suggest that features of the speech signal such as the frequency and bandwidth of formants, vocal tract size, fundamental frequency, F0, and spectral envelope shape contribute substantially to a listener’s ability to recognize, identify, and discriminate individual voices [1, 2]. However, many of these acoustic features vary within an individual speaker as well as across speakers when the level of vocal effort varies [3, 4, 5]. For example, Pickett found frequency-dependent changes in level within an individual speaker when going from conversational speech to shouted speech that resulted in overall changes in the spectral tilt of the speech signal. Other studies have shown that values of F0 are much less differentiated across speakers for shouted speech than for conversational speech and are comparable to the frequency of F1 found in conversational speech [4]. Shouted speech has also been shown to lead to a lengthening of syllables, reduced utterance speed, and less variability in intonation both within and across speakers. It has been suggested that such convergence of acoustical features makes discrimination between individual talkers more difficult [4]. Similarly, the properties of whispered speech differ substantially from conversational and shouted speech, most notably in the lack of voicing. The absence of F0 reduces the differences in acoustic features across talkers, and reduces the similarity across vocal effort within talkers. In addition, the formant structure of whispered vowels is distorted due to restricted flexibility of the vocal tract, and insufficient air pressure may obscure the production of consonant sounds. These distortions may disrupt the identification of individual speakers in unvoiced speech.

The experiment described in this paper was designed to examine the influence that vocal effort has on speaker identification by human listeners. The experiment had two objectives. The first objective was to determine the impact that vocal effort level has on a listener’s ability to learn to identify the individual talkers in a series of speech stimuli. In other words, whether it is easier to identify whispering talkers, conversational-level talkers, or shouting talkers. The second objective was to determine the extent to which learning to identify a set of talkers speaking at one level of vocal effort allows listeners to identify the same set of talkers when they are speaking at a different level of vocal effort. In other words, whether the ability to identify conversational-level talkers implies the ability to identify the same talkers when they are whispering or shouting. The next section describes the stimuli and procedures used in this experiment.

2. Methods

2.1. Speech stimuli

Eight adult talkers (four male, four female) were used to record the speech signals used in the experiment. The measurements were made with a 1” pressure microphone (B&K 5935) located 1 m from the talkers’ mouths in a large anechoic chamber. The talkers were first asked to repeat three short utterances in their quietest whispered voice: “warning;” “over here;” and “threat.” The phrases were recorded digitally, and the talkers were asked to repeat the three utterances in a slightly louder whispered voice, and again in an even louder whispered voice, until they indicated that it was no longer possible to whisper any louder. Then the talkers were asked to repeat the three utterances in a conversational voice, and to continue repeating the utterances in a louder and louder conversational voice until they felt they were unable to speak any louder without shouting. Finally, the talkers were asked to repeat the utterances in their quietest shouted voice and to gradually raise their voice until they were shouting as loudly as possible. The number of...
levels of whispered, conversational, and shouted speech varied considerably across the talkers, but the range of output levels (measured by the RMS power of the speech signals at 1 m) consistently ranged from \( \approx 35 \) dB SPL for the quietest whispers to \( \approx 100 \) dB SPL for the loudest shouts.

A subset of these speech signals was selected for use in each testing and training mode of this experiment. In the whispered speech modes, the speech stimuli were selected from the three least intense whispered samples of each utterance spoken by each of the eight talkers. In the conversational speech modes, the speech stimuli were selected from the three samples of each utterance produced closest to 65 dB SPL by each talker. In the shouted speech modes, the speech samples were selected from the three most intense shouted samples of each utterance produced by each talker. Note that all of the stimuli were scaled to the same overall RMS level (approximately 65 dB SPL) before they were presented to the listeners in the experiment.

2.2. Procedure

A total of 30 paid volunteer listeners were recruited for the experiment. These volunteers all had normal hearing, and none had any prior exposure to the talkers used to record the stimuli. Each volunteer was randomly assigned to participate in two different experimental sessions—one where they were asked to identify the four male talkers used in the experiment, and one where they were asked to identify the four female talkers. Within each of these sessions, they were randomly assigned to be trained on one of the three different levels of vocal effort (whispered, conversational, or shouted). Thus, a total of 60 experimental sessions were collected in 6 different training modes (three levels of vocal effort and two talker sexes).

The experimental sessions were conducted with the listeners sitting at the console of a control computer in a sound-deadened listening booth. A flow diagram of a typical experimental session is shown in Figure 1. Each session started with an initial training condition where speech stimuli were randomly selected from the 36 possible speech samples for that category of speech (3 utterances [“warning,” “over here,” and “threat”] * 3 levels * 4 talkers) and presented to the listener over headphones (Sennheiser HD-520). The listeners were asked to identify which of the four talkers was speaking in each stimulus presentation, and to respond by using the mouse to select one of the numbers 1-4 on the CRT of the control computer. Then they were provided feedback about the actual talker number used in the stimulus. This procedure was repeated until either A) the listener met the training criterion of the experiment by responding correctly in 90% of the most recent 50 trials of the experiment; or B) the listener participated in 500 consecutive trials without meeting the training criterion. In the latter case, the listeners were excused for the day and the same training procedure was repeated on the next day with a 10% reduction in the training criterion. For example, listeners who failed to meet the criterion on the first day would train for up to 500 trials on a second day with an 80% criterion. Table 1 shows the number of training sessions required for the listeners in each of the six training modes of the experiment.

Once the training criterion was met, the listeners proceeded into a series of four blocks of data collection. In each of these blocks, they heard a total of 72 stimuli consisting of two replications of each of the 3 utterances, 3 levels, and 4 talkers available for that particular mode of speech. As in the initial training condition, the listeners were asked to respond by identifying which of the four talkers was speaking in each stimulus presentation. However, they were given no feedback about the correct identity of the talker in these testing blocks. The first and last testing blocks used the same level of vocal effort used in the initial training condition. The second and third testing blocks used the two levels of vocal effort not used in the initial training condition. Thus, in the example shown in Figure 1, the listeners trained on conversational speech and were tested on conversational speech in the first and fourth data collection blocks, on whispered speech in the second data collection block, and on shouted speech in the third data collection block. Between each pair of data collection blocks, the listeners participated in an additional training condition with feedback until they met the same training criterion achieved in the initial training phase of the session. Note that testing order of the second and third data collection blocks was selected randomly for each listener, and that in all cases the four data collection blocks were completed on the same day that the listeners first met criterion in the initial training phase of the session.

Upon completing the first experimental session, the listeners were asked to return on another day to begin initial training in a second experimental session with different-sex talkers. Upon completion of the second session, they were paid and released from the experiment.

3. Results

3.1. Training time

One measure of the overall difficulty of speaker identification at each level of vocal effort is the number of trials required to reach the criterion level of performance in the initial training phase of each experimental session. Table 1 summarizes this data for the six training modes used in the experiment. By far, the most difficult training modes were the ones that used whispered speech. None of the listeners were able to meet the 90% criterion in these training modes, and for both the male and female talkers there was one listener who was unable to meet the training criterion until the fourth day of training when the criterion was reduced to 60%. Overall, it took about three times as many trials to meet the criterion with the whispered speech than with the conversational speech. These data suggest that the lack of voicing in whispered speech substantially impairs a listener’s ability to distinguish between speech waveforms spoken by different talkers.

The easiest training modes were the ones that used shouted speech. Only one listener was unable to meet the training criterion on the first day with shouted speech, and the overall average number of trials needed to reach the criterion was about half the number needed to reach criterion with conversational speech. Thus, it appears that it was substantially easier to identify the talkers in this experiment when they were shouting than when they were speaking at a conversational level.
Table 1: Training time required to meet criterion for each of the training modes used in the experiment. The numbers in the table represent the number of listeners who completed the initial training in the number of days (and at the criterion level) specified at the top of each column. The numbers in parentheses represent the average number of trials required to meet criterion for the listeners who completed the test in the specified number of days. The last column represents the overall average number of trials needed to meet criterion across all the listeners in that training mode.

<table>
<thead>
<tr>
<th>Training Mode</th>
<th>Male Talkers</th>
<th>Mean Trials</th>
<th>Female Talkers</th>
<th>Mean Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whispered</td>
<td>1d(90%)</td>
<td>2d(80%)</td>
<td>3d(70%)</td>
<td>4d(60%)</td>
</tr>
<tr>
<td>Conversational</td>
<td>7 (191)</td>
<td>2 (748)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Shouted</td>
<td>9 (153)</td>
<td>1 (550)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

3.2. Transfer of training across vocal effort levels

In general, performance was poor when listeners who were trained on speech at one level of vocal effort were tested on speech at a different level of vocal effort. Tables 2 and 3 show the overall percentage of correct identifications for each combination of training and testing modes in the experiment. In conditions where listeners were trained and tested on the same type of speech (the bold numbers along the diagonal), performance was generally good. Note, however, that performance was substantially lower for whispered speech than for the other types of speech (see Table 1).

When the listeners were trained and tested on different types of speech, performance decreased substantially. This is shown in more detail in Figure 2, which shows performance in each training and testing condition in terms of bits of information transfer (IT). IT is a better overall measure of performance than percent correct because it takes into account the effects of guessing and the differences in performance across the different talkers used in the experiment. The IT values in the figure were calculated directly from the confusion matrix in each condition [6]. The percentages shown above the bars in the figure represent the information transfer in each testing condition as a percentage of the information transfer in the testing condition with the same vocal effort level used in the training mode. For example, in the whispered training mode with male talkers, the information transfer was 0.73 bits in the whispered testing mode, 0.25 bits in the conversational testing mode (34% of 0.73), and 0.11 in the shouted testing mode (15% of 0.73).

In most cases, only a small portion of the information transfer achieved in the testing mode with the same vocal effort as the training mode was maintained when the listeners were tested on speech produced at a different level of vocal effort. In general, a higher level of performance was maintained across the whispered and conversational modes (22%, on average) and across the conversational and shouted modes (31%) than across the whispered and shouted modes (11%), which is not surprising because the whispered and shouted modes were much further apart in terms of overall vocal effort. The greatest transfer of training across different vocal effort levels (52%) occurred when the listeners were trained on male conversational speech and tested on male shouted speech. It is not clear why this transfer was so large relative to the corresponding condition with female speech, where the information transfer was only 14%. Training times were similar for the male and female conversational speech (Table 1), so it doesn’t appear that the listeners were any better trained on the male conversational talkers than on the female conversational talkers. Information transfer in the symmetric condition from shouted speech to conversational speech was almost identical for the male and female talkers (27% vs. 29%), so it doesn’t seem likely that the conversational and shouted speech samples were somehow “more
than conversational-level talkers [4]. One possible explanation is that the listeners were basing their judgments, in part, on differences in the apparent production levels of the different talkers. Because the talkers varied in their ability to produce loud speech, there were some variations in output level across the different shouting talkers used in the experiment. The male Talker 2 was consistently at a higher production level than the other shouting male talkers (shown by the diamonds in the upper left of the top panel of Figure 3), and the female Talker 1 was consistently at a lower production level than the other shouting female talkers (shown by the triangles in the upper left of the bottom panel of Figure 3). In the case of the male talker, there is some evidence that this discrepancy in production level had an impact on performance—the listeners were able to identify the shouted speech of Talker 2 nearly 100% of the time. However, there is no indication that production level had any impact on performance for the female talker, and overall training time was actually lower for the female talkers than the male talkers. Thus, it does not appear that differences in production level alone can account for the fast training times with shouted speech. Rather, it seems that one or more of the cues that the listeners were using to discriminate between the different talkers was exaggerated, rather than diminished, in the high-level speech. Further research is needed to determine which acoustic cues listeners use when they are identifying shouting talkers.

The results also make it clear that training on the vocal characteristics of different talkers at one level vocal of effort is not sufficient to allow human listeners to reliably identify those talkers when they are speaking at a different level of vocal effort. This result has implications in the design of automatic speech identification systems. In complex pattern recognition tasks like the speaker identification task in this experiment, one would expect human listeners to have substantial advantages over automatic speaker identification systems. The fact that listeners were unable to reliably perform this task suggests that automatic systems are likely to have substantial difficulty identifying talkers who are speaking at a different vocal effort level than the one used to train the system. Designers of automatic recognition systems should be cognizant of this potential problem and address it either by training their algorithms at a variety of different vocal effort levels or finding some way to control the vocal effort level of each test utterance.

5. References