MACHINE RECOGNITION VS HUMAN RECOGNITION OF VOICES

Stanley J. Wenndt

Air Force Research Laboratory
Rome, NY 13440, USA
Stanley.Wenndt@rl.af.mil

Ronald L. Mitchell

Clarkson University
Potsdam, NY 13699, USA
mitcherl@clarkson.edu

ABSTRACT

While automated speaker recognition by machines can be quite good as seen in NIST Speaker Recognition Evaluations, performance can still suffer when the environmental conditions, emotions, or recording quality changes. This research examines how robust humans are compared to machine recognition for changing environments. Several data conditions including short sentences, frequency selective noise, and time-reversed speech are used to test the robustness of both humans and machine algorithms. Statistical significance tests were completed and, for most conditions, human were more robust.

Index Terms**—**Speaker Familiarity, Human Voice Recognition, Robust Speaker Identification

1. INTRODUCTION

For speaker recognition, there may be many clues that convey the speaker's identity such as word choice, accent, dialect, gender, etc. Speaker recognition can be separated into two categories of speaker identification and speaker verification. Speaker identification involves identifying the person who is speaking. Speaker verification involves comparing two voices and deciding if they are from the same person or not. This paper takes a closer look at the speaker identification task by humans using a closed set of speakers that are familiar to them (i.e., co-workers).

The goal of this research is to not only learn how robust humans are or are not for speaker identification in changing environments, but also to learn what cues may play an important role. The assumption at the start of this research is that humans are more robust for familiar speaker identification in changing environments.

2. LITERATURE REVIEW

Speaker identification and speaker verification by listeners has been studied for a while. For example, an early reference [1] is from 1966. 10 male speakers were recorded and their voices were presented to 16 listeners who were familiar with the speakers. For short sentences (average length of 2.4 seconds), the accuracy was 98%. For speaker recognition tests, accuracy is defined as whether or not the listener identified the correct speaker. It doesn't take into account how many words were correctly recognized. The accuracy of speaker recognition for disyllables was 87%. For monosyllables, it was 81%, consonant-vowel excerpts were 63%, and vowel excerpts were 56%. Thus, they demonstrated that the identification performance decreased as the number of phonemes decreased.

In [2], the authors used 45 voices from famous people such as Johnny Carson, Bob Hope, Lawrence Welk, etc. The research looked at how well subjects could recognize the voice when it's played the normal way (forward) and when it's played backwards (time-reversed). Recognition rates for the forward rate were about 71%. For the backward voice, the general trend was that it was harder to do voice recognition compared to the normal audio (about 12% decrement in performance). However, some voices could be recognized almost equally as well when played backwards, while other voices suffered a very large drop in performance when played backwards.

In [3], the authors used the same voices of famous people and examined the effect of time-altered audio by either expanding or compressing the audio. The results were somewhat similar in that some voices could still be easily recognized when time-altered while others could not. But, they concluded that they were not able to predict which voices could still be recognized after the rate change or which traits were key for recognition.

The research in [4] had a relatively large set of 24 speakers and 24 listeners. The focus was to see how audio coding (in this case linear predictive coding) affected the performance of speaker familiarity. Not surprisingly, the performance was lower for the coded audio compared to the original audio, but a lot of speaker information was still retained. For the original data, the recognition rate was 88% and for the coded speech it was 68%.

Other research [9] examined speaker recognition where there were 5 female and 5 male speakers to identify. The error rate for high quality speech was 23% but for telephone speech it was 48% and for LPC speech, it was 46%. For this research, the voices were taken from a larger collection and may not have been familiar to the listeners. A reference sample was provided to allow comparison to the stimuli.

The challenge with familiar speaker recognition experiments is that it is difficult to get a large group of people who are familiar with each other. Thus, experiments tend to stay at about 10-15 speakers. A different approach is speaker verification which is another human listening task. For this task, the subject is presented with two voices and has to decide if it's the same or different speaker. Usually, the voices that are presented are not familiar to the listener. For this type of experiment, one can get a much larger set of human responses.

In [5], the research compared speaker verification of human listeners to that of machines. After training on a speaker, listeners were tested on 21 test samples where the test sample duration was three seconds. There were 65 listeners and 144 target speakers. In total, the listeners yielded 48,972 decisions where the listeners had to decide if the speakers from the training data and the 3-second test sample were the same or different. The human listeners were comparable to the machine algorithms, but humans were more robust to changing environments. An earlier reference on speaker verification using customers, imposters, and mimics

One article took a deeper look at where speaker identification and speaker verification occurs in the brain. The article [6] showed that the discrimination of unfamiliar voices (speaker verification) compared to the recognition of familiar voices (speaker identification) were distinct processes that occurred in different parts of the brain. The research used patients with either left brain damage or right brain damage. Having a lesion in the left hemisphere of the brain caused a deficiency in the discrimination scores (speaker verification).

3. EXPERIMENTAL GOALS

The goal of this research was to examine how robust humans are in recognizing familiar voices and compare that to machine performance. Additionally, the goal was to glean information about what cues humans may be using. The long-term goal is to use this information to foster additional research for feature development.

There were 17 participants in this study which included 3 females and 14 males (USAF IRB, Protocol F-WR-2010-0028-H). A pure tone test was used to test their hearing. Of the 17 participants, 11 participants had normal hearing. The other 6 participants failed at least one frequency. The frequencies tested were 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. For this research, normal hearing is defined as 25 dB above the ANSI hearing threshold. The results are reported in two groups: those with Normal Hearing (NH) and those with a Hearing Deficit (HD).

For this research, the listeners were tasked with identifying familiar voices. From the MARP corpus [7], there were 25 voices (20 males and 5 females) that were familiar to the listeners in this research. The audio data collected under the MARP corpus was designed to test various factors that affect speaker identification using machine algorithms. Some of these factors included aging (the data was collected over three years), speaking style, duration of the audio, and intonation. For the speaking style, there was read data, whispered data, conversational data, and "naturally spoken" data.

Table 1: List of Short Sentences

Sentence	Sentence
1	Let's go skiing today.
2	We'll be leaving early tonight.
3	You're going to go with them.
4	It's time to go now.
5	We could get a drink.
6	I need some coffee now.
7	She was home too late.
8	He broke his lower leg.
9	We need to be careful.
10	He heard the movie was great.

The naturally spoken data was designed to elicit daily communication words and style, such as "We could get a drink" where the words are spoken easily and rapidly (see Table 1 for the 10 sentences that were used). The 10 sentences were designed to be short in duration (about 1-2 seconds), have sufficient phonetic coverage, and to be spoken naturally. While the MARP data was not originally designed with human listening experiments in mind, there were some very strong attributes to it.

4. EXPERIMENTAL SETUP

After the pure tone test, a familiarization phase was completed to allow the listeners to become accustomed to the experimental setup and the tasks at hand. Data other than the 10 short sentences were used for the familiarization phase. For this phase only, the listeners were allowed to repeat the audio and given feedback as to the correct answer. As stated before, there were 25 voices for playback. When the listener would hear a voice, they would be required to choose from a drop-down list of all 25 voices.

After the training phase, there was a baseline experiment with clean stimuli (no additive noise). This was followed by several experiments with various types of noise degradation. This research used additive speech-shaped noise (LTASS) to degrade particular frequency regions of the speech signal [8]. This way, the signal will still sound natural and the performance of listeners could be tied directly to the degradation of particular frequencies. If the performance decreases when a set of frequencies are masked by an interfering signal, it would indicate that frequency range was important. Additionally, time reversed speech was presented to the listeners. For all noise scenarios, no audio examples were provided to allow them to become accustomed to the noise. All listening experiments were conducted in a soundproof booth.

Table 2 lists the session number, the noise location, and the noise level. A -20 dB signal-to-noise ratio for each noise region is chosen to mask that particular region. Other techniques such as bandfiltering could be used to exclude frequency regions but would alter the naturalness of the audio. As seen in Table 2, Sessions 1-5 used new sentences to prevent a learning curve from hearing the same stimuli over and over. Session 6 was identical to Session 1 to see if there was a learning curve.

Table 2: Session Number and Corresponding Noise.

Session	Noise Location	SNR Level	Sentence #
1	Clean-1	N/A	1, 2
2	0-1000 Hz	-20 dB	3, 4
3	1000-2000 Hz	-20 dB	5, 6
4	2000-3000 Hz	-20 dB	7, 8
5	3000-4000 Hz	-20 dB	9, 10
6	Clean-2	N/A	1, 2
7	Time-reversed	N/A	1,2
8	Whispered	N/A	N/A

5. EXPERIMENTAL RESULTS

This section presents the speaker identification results for the human listeners and the machine algorithms. Then, a statistical comparison is completed to see if the differences are significant.

5.1. Performance for Human Listeners

Figure 1 shows the results for Sessions 1-6 for both the NH listeners and the HD listeners. Remember that there are 25 voices and for every session, each voice is presented twice. The sentences were downsampled to 8000 Hz for all experiments. Thus, a 90% correct means that out of 50 voice presentations, the listener identified the correct speaker 45 times. It was always a forced choice decision. Note that in

Figure 1, the two lines follow the same general trend except the HD group seemed to have a larger, broader dip. Not surprising, all speakers do the best with no additive noise. Any additive noise causes a drop in performance for both groups.

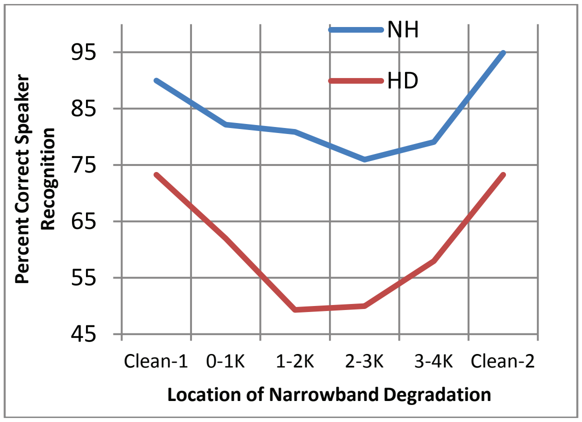

Figure 1: Experimental Results of Table 1.

5.2. Performance for Machine Algorithms

For the machine algorithms, various classifiers, features, settings, and combination of these things were tried. In the end, a GMM-UBM classifier with multiple features was used. The features used were mel-frequency cepstral coefficients, perceptual linear prediction, and linear prediction cepstral coefficients. Additionally, RASTA filtering, voice activity detection, Gaussian super vector, and cepstral means and variance normalization was used. The number of mixtures for the universal background model (UBM) was 256. The goal was not to optimize the test scores for this particular experimental setup, but to use a classifier that has been shown to be robust across many different conditions. The UBM was built with over 900 unique speakers and about 20 hours of audio data (prior to silence removal).

There were two techniques for the speaker identification scores by machines. The first technique trained the UBM and speaker models with only clean data (i.e., no additive noise). Thus, if the test case was LTASS noise at -20 dB in the 1000-2000 Hz frequency range, there was a strong cross-condition between clean training data and noisy test data. This is similar to the human listeners in that the listeners were not given any exposure on the noise conditions. However, the listeners can adapt to noise and can draw on past experiences. The results for this scenario are presented in Figure 2. Remember that there are 25 voices and for every session, each voice is presented twice.

In Figure 2, the results of the NH group are repeated here for convenience. For the Clean-2 case, the machine algorithm result does not change from the Clean-1 case since the algorithm is deterministic.

In order to eliminate cross-conditions between the training and testing data, a second set of experiments was run where all the data had the same noise condition. For example, if the LTASS noise condition was -20 dB in the 1000-2000 Hz, then the UBM data, the training data, and the test data would all have the same LTASS noise condition. The results for this scenario are presented in Figure 3. Once again, the NH results are presented for convenience.

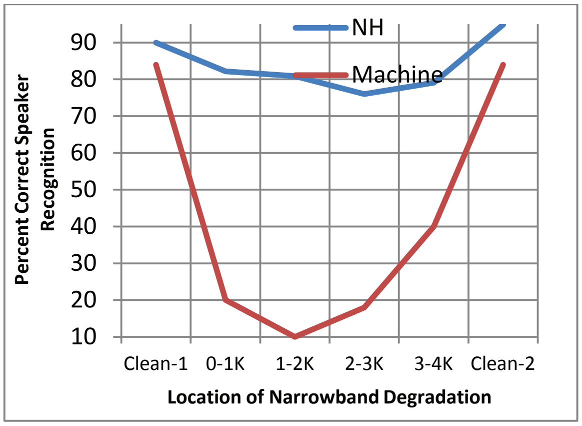

Figure 2: Machine Algorithms Results with Cross-Conditions

5.3. Statistical Analysis

The previous two sections examined the results of the human listeners compared to machine algorithms. This section does a statistical analysis for the results in Figure 2 and Figure 3 to compare the mean or median of the NH group to the machine algorithm results.

Beginning with the results of Figure 2, it seems obvious that the results are statistically different. However, it is still important to outline and step through the process of reaching this determination. The first step is to test if the NH distribution is normal or not for each noise condition. A Jarque-Bera test was used to test if each distribution is a normal distribution with a significance of 0.05. The null hypothesis of the Jarque-Bera test is that the distribution is a normal distribution with unknown mean and variance. The Jarque-Bera test indicated that only two noise conditions of LTASS noise in the 2000-3000 Hz and 3000-4000 Hz range pass the Jarque-Bera test.

For these two conditions, the next step is to use the T-test. The null hypothesis is that the mean of the NH group is statistically greater than the result of the machine algorithm. A right tailed test is used which tests the alternative hypothesis. For both conditions, the alternative hypothesis fails (i.e., the mean of the NH distribution is statistically greater than the machine result).

The other conditions which failed the Jarque-Bera test, the next step is to use the signrank test. The null hypothesis is that the NH distribution is a symmetric distribution with a median equal to the machine result. Only one condition, the Clean-1 condition, is not statistically different than the machine results. For this condition, the median of the NH distribution was 94% and the machine result was 84%.

For Figure 3, the same steps are followed. Once again, for the Clean-1, the NH result and the machine result are not statistically different. Additionally, the noise case of -20 dB, 0-1000 Hz LTASS yielded this same result. For the other four conditions, the human scores were statistically different that the machine scores.

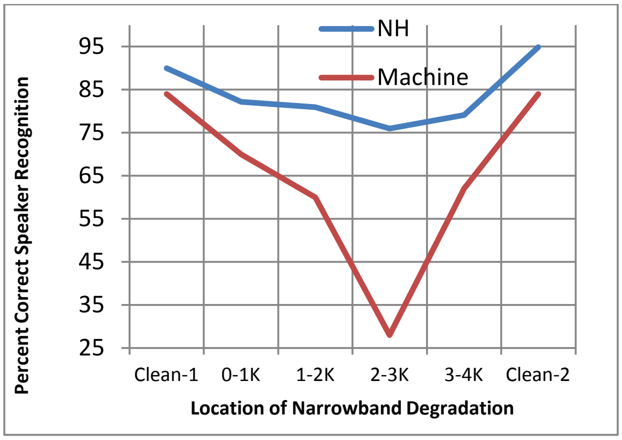

Figure 3: Machine Algorithm Results for Matched-Conditions

6. TIME-REVERSED SPEECH

Table 3 shows the results for the time-reversed speech data. For the time-reversed speech, Sentence 1 and 2 of Table 1 were used. The temporal cues such as voice onset time, rising or falling F_0, or phonetic order were distorted; however, other information such as speaking rate, F_0, and formants are still present.

For the results in Table 3, the Machine-Cross condition means that the UBM and the speaker models were developed with normal speech, but the testing was done with time-reversed audio. Thus, there is a cross-condition between training and testing. For the Machine-Same condition, the UBM, the speaker models, and the testing were all done with time-reversed audio which alleviated a cross-condition.

Comparing the NH group to the Machine-Cross condition, the results are similar. The machine algorithm did about 5% better, but there is no statistical difference. For the Machine-Same condition, the machine algorithm was 23% better and is statistically different to 0.05 using the Jarque-Bera test and the T-test.

Similarly to the conclusions of [2] and [3], some voices could still be easily recognized when time-altered while others could not. There were a few voices that tended to be easily recognized in the conditions of Table 2. Two of these voices were still recognized at a high rate for the time-reversed scenario.

Table 3: Results for Time-Reversed Speech

Group	Time-Reversed
NH	56.9
Machine-Cross	62.0
Machine-Same	80.0

7. SUMMARY

The goal of this research was to see if human listeners were more robust for the speaker identification task compared to a machine algorithm. Even when the machine algorithm is trained to match the noise environment, most results showed that humans were more robust. The strength of the human listeners were especially evident for the challenging case of noise case in the 2000-3000 Hz frequency range. This frequency range can contain the 2nd, 3rd, and/or the 4th formant. Thus, there is a lot of lexical and speaker cues that may be masked. For both NH listeners and the machine, this was the most challenging noise scenario. Perhaps this suggests some future research for feature development.

Additional analysis is looking at various factors that may impact a listener's ability to identify a person's identity. For example, the amount of voiced (or unvoiced) speech was examined to see if there was a correlation with how easily a speaker's voice was recognized. Unfortunately, the amount of voiced (or unvoiced) speech did not correlate strongly with how easily a speaker's voice was recognized. Other factors such as fundamental pitch, formant locations, pitch shimmer, pitch jitter, and other modulation measures are being examined.

The original goal of this effort was to discover which frequency bands are most important for the familiar speaker recognition task. While there is some research in the literature that looked at how well listeners could identify familiar speakers, the author did not find research that looked at what frequency information was important for speaker identification. This research was a cursory look and requires more listening experiments with better randomization of stimuli and phonetic consideration.

8. REFERENCES

[1] Bricker, P., Pruzansky, S., "Effects of Stimulus Content and Duration on Talker Identification," *Journal of the Acoustical Society of America*, 1966, Vol. 40, pp. 1441-1449.

[2] Van Lancker,D., Kreiman, J., Emmorey, K. "Familiar Voice Recognition: Patterns and Parameters Part I: Recognition of Backward Voices, "Journal of Phonetics, 1985, Vol. 13, pp. 19-38.

[3] Van Lancker,D., Kreiman, J., Wickens, T. "Familiar Voice Recognition: Patterns and Parameters Part II: Recognition of Rate-Altered Voices, "Journal of Phonetics, 1985, Vol. 13, pp. 39-52.

[4] Schmidt-Nielsen, T., Stern, F., "Identification of Known Voices as a Function of Familiarity and Narrow-Band Coding,", *Journal of the Acoustical Society of America*, Feb 1985, Vol. 77, No. 2, pp. 658-663.

[5] Crystal, A., Schimdt-Nielsen T., "Speaker Recognition by Human Listeners: Experiments Comparing Human and Machine Performance Using the NIST 1998 Speaker Evaluation Data," *Digital Signal Processing*, 2000, Vol. 10, pp. 249-266.

[6] Van Lancker,D., Kreiman, J., "Voice Discrimination and Recognition are Separate Abilities," *Neuropsychologia,I* 1987, Vol. 25, No. 5, pp. 829-834.

[7] Lawson, A., et. Al., "The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design, and Initial Findings, Interspeech 2009, pp. 1811-1814.

[8] Byrne, D., et. al., "An International Comparison of Long-Term Average Speech Spectra," *Journal of the Acoustical Society of America*, 1994, Vol. 96, No. 4, pp. 2108-2120.

[9] Papamichalis, P., Doddington, G.. "A Speaker Recognizability Test,". Proceedings ICASSP, 1984.