

Testing the Intelligibility of Corrupted Speech with an Automated Speech Recognition System

William T. HICKS, Brett Y. SMOLENSKI, Robert E. YANTORNO
Electrical & Computer Engineering Department
College of Engineering, 12th & Norris Streets
Temple University, Philadelphia, PA 19122-6077

and

Norman E. SHAW
Triton PCS
1100 Cassatt Road
Berwyn, PA 19315

ABSTRACT

Co-channel speech is defined as two people talking at the same time. Much research has been and is continuing to be done in the area of separating co-channel speech into the original utterances of the two speakers. This is very important in the area of automated speech recognition, where the current state of technology is not nearly as accurate as human listeners when the speech is co-channel. As part of this effort, an objective method to compare the success of various reconstruction algorithms is desired. This research defines one such method and gives the results of testing the technique on stationary noise and co-channel speech. It was found that the software used in this method works better on co-channel speech than on speech in stationary noise, and that the success of the software to work at various levels of corruption is comparable to or better than previously reported. The success rate varied from 97% for single speaker speech to 19% for co-channel speech with a -6 dB TIR.

Keywords: Co-channel speech, speech reconstruction, speech intelligibility, usable speech, SPHINX.

INTRODUCTION

Since 1998, the Speech Processing Laboratory of the Electrical and Computer Engineering Department of Temple University has been conducting research in detecting and extracting usable speech from co-channel speech. Usable speech (as defined here) is that speech from the co-channel speech that can be used for some process, like speaker identification.

One of the goals of the research project of the Lab (see Figure 1 below) is to extract usable segments, separate them into two groups (speaker #A1 and speaker #A2) and then fill in the empty segments of each speaker's speech to form the original utterance of each speaker. Figure 1 shows the main parts of this process. Previous work has been to find methods of identifying which segments of co-channel speech are usable and methods of identifying which usable segments belong to which speaker.

The first portion of that research was in the detection of usable speech. So far three methods have been developed for finding usable speech segments in co-channel speech.

One method for finding usable speech segments in co-channel speech is called Spectral Autocorrelation Ratio (SAR), later called Spectral Autocorrelation Peak Valley Ratio (SAPVR) [18]. This method was later refined and was described as the SAPVR using LPC residual [2]. The SAPVR residual measure takes the autocorrelation of the spectrum of a segment of speech, and then compares the level of the first peak to the level of following valley's. This results in the SAPVR of a signal that has a much flatter and more periodic spectrum than the original signal. It was found, that by selecting an appropriate threshold of the ratio of the peak-to-valleys, usable segments of the co-channel speech could be identified. The usable segments had a higher ratio of peak-to-valleys than the non-usable segments. The usability criteria was defined in terms of using the segments in a speaker identification system that used LPC cepstral coefficient or LPC residual methods [12].

The second measure is the Adjacent Pitch Period Comparison (APPC) [11]. This measure uses the time domain and compares adjacent pitch periods in a segment of speech to see if there is a good comparison between them. It relies on that fact that usable speech is single speaker, and therefore the pitch periods are repeated. Again, it was used to find usable speech segments for speaker identification. Speaker identification was found to require the target speaker to be at least 20 dB above the interfering speaker for 80% success [9].

The third measure was the Peak Difference of Autocorrelation of Wavelet Transform (PDAWT) [8]. Using this method one first finds if the segment is voiced, then takes the discrete wavelet transform (DWT) of the signal. This is followed by taking the autocorrelation of the first half of the DWT and looking at the peaks to find if the segment is rich in one signal and its harmonics. Again a 20 dB TIR threshold was selected for defining usable segments.

In another research project, the Lab personnel have looked to see if the usable speech methods can be combined or "fused" to

obtain a better overall method. This "fusion method" has been successful when there was sufficient independent information in the various methods to get a better overall measure. The first fusion method used independent components and nonlinear estimation [16]. First, independent component analysis was performed to eliminate any redundancy between the signals, then the non-linear minimum mean square error estimate was used. This gave an improvement in both finding usable segments and reducing false alarms. This research was later followed by a method that used quadratic discriminant functions for the optimal Bayesian classifier, to fuse two previous measures [15]. Again improvement was found over using either measure separately.

Another part of the research in the Lab is to take all the information available and reconstruct the speech of each speaker. As part of the speech lab's effort, an objective method to measure the success of various reconstruction methods is desired. This research reported here is the results of defining one such method, and calibrating the method, so it can be used in further research. The method is a way of performing relative comparisons of intelligibility between speech files, using speech recognition software.

SUBJECTIVE VERSUS OBJECTIVE TESTING, MAN VERSUS MACHINE

Human listeners can tolerate a much higher level of interference in corrupted speech than that obtained from using speech recognition systems, while still obtaining understanding. "Human speakers are remarkably good at understanding speech in noise backgrounds. Psychophysical data suggests that listeners group features in complex 'auditory scenes' into streams which allow selective listening. [13]." In work done on co-channel speech [6], one of the findings was that speech is highly intelligible at 0 dB, but unintelligible at -6 dB (TIR), for closely spaced pitch. They found that on a frame basis, 0 dB is a threshold for human listeners that marks a boundary between intelligibility and unintelligibility success or failure. They also obtained data on unprocessed co-channel speech, with the following results. Using one test method they found that at 0 dB TIR: 88% correct, using a different test method they found that at -6 dB TIR: 73.3% correct, and at -12 dB TIR: 53.8% correct. As part of a project [17], the DRT (Diagnostic Rhyme Test [1]) was used in comparing two sets of data, one with no interference and one with 0 dB MNRU (Modulated Noise Reference Unit [7], which adds a Gaussian white noise to the short term signal level rather than the average signal level). The results were (for American speakers and American listeners) 96.5 percent correct for no noise and 72.3 % correct for the noise case.

For automated speech processing, [4] describes how background noise degrades automatic speech recognition systems. Another work, [14], found that "For the human intelligibility problem, the desired talker is the weaker of the two signals with voice-to-voice power ratios (Power desired / Power interference), or VVRs, as low as -18dB. For automatic speech and speaker recognition applications, the desired talker is the stronger of the two signals, with VVRs as low as 5 dB." Note that this gives a 23 dB difference between human and machine listeners. The tests described above are summarized in table 1 below.

As can be seen from these examples, the use of automated speech processing at recognizing corrupted speech is not as effective as human listeners. However, for many applications, automated speech recognition (ASR) is required or is the most efficient method available, ASR can be used to compare intelligibility between various levels of interference. Because the final goal of this research will be for automated testing of various methods to reconstruct co-channel speech, an automated speech recognition method was used.

EXPERIMENTAL METHODS

In order to keep the task controllable, and the dictionary manageable, the TIDIGITS [10] database was selected. This contains the ten digits, including two versions of zero, for a total of eleven words. In casual human listening tests, the words zero and oh were confused, so the oh was removed in case future tests use human listeners, leaving only ten words. Tests were run on various multiple word and single word utterances of digits, spoken by various speakers of both sexes. A test of this database with uncorrupted utterances provided a base-line of the best performance.

In order to compare the automated method to previous work done with human listeners and noise interference, some tests were run on the TIDIGIT (target utterances) database with just Gaussian white noise added. These tests were run using an untrained speech recognition system.

We also performed a small test with both stationary Gaussian noise as above, and non-stationary modulated noise as described previously. This second noise test used a trained speech recognition system.

A second database was selected to use as the corrupting utterances. The TIMIT [5] database was chosen for its rich range of speakers and utterances.

The target utterances were paired with the closest longer file of the appropriate sex from the interfering utterances. As the interfering files were longer than the target files, the interfering files were truncated to the same length as the target files. The interfering file truncation was linearly tapered at the end from no attenuation to full attenuation.

The energy level of each file was measured after padding and included silent segments, and the average energy for the utterance was calculated. The TIR (Target to Interferer Ratio) was then produced by adjusting the interfering file level to the appropriate dB relative to the target file. These tests were done using an untrained speech recognition system.

RESULTS

Gaussian White Noise

In order to compare the automated method to previous work done with human listeners, some tests were run on the TIDIGIT database using only Gaussian white noise. Noise of equal energy was added to all parts of the utterance. The tests were run at various SNR levels. The number of correct hits was counted by counting each correctly identified digit. As long as the results had a digit in the same sequence as the input, a success was recorded. Extra digits in the results, which

represented a small portion of the samples, were not considered in the tabulation of accuracy. Alternatively, the extra digits could have been listed as false alarms. The results of the tests are listed in Table 2 below. As was expected, the more the noise the less the percent correct. Note that by -6 dB SNR the system has stopped recognizing that there was even an utterance.

Another experiment was conducted using the "Dragon Systems Naturally Speaking" speech recognition system and training it with data from a data base of digits made by ourselves. Two types of tests were run. One was with the noise level constant and stationary, as was done in our first test. The second was with the noise level calculated on a frame by frame basis, similar to the modulated noise tests described previously. The results of these tests are listed in Table 3 below. As expected the higher the level of interference the lower the ability of the speech recognition software to correctly find the digits. Also note how the degrading of men is slower than that of women.

Co-channel interference

Tests were run on the TIDIGIT database with TIMIT utterances added. The tests were run at various TIR levels. The number of correct hits was counted by counting each correctly identified digit. As long as the results had a digit in the same sequence as the input, a success was recorded. Extra digits in the results, which were a small portion, were not considered in the tabulation is of percent correct. The results of the tests are listed in Table 4 below. Again the more the interference the less the percent correct. While men are better correctly detected than women, it is not as pronounced as the previous cases.

CONCLUSIONS

While the results of our research are useful, it is also instructive to compare them to previous work in this area. Earlier work did not follow a similar testing mythology as was used in our investigation, however, some conclusions can be drawn. The previous work with human listeners [17] in one test shows 72 % correct with 0 dB MNRU noise. Another test [6] showed 73 % correct for -6 dB TIR (which is an average level of co-channel interference). Therefore, the previous two tests show a similar success rate for 0 dB MNRU noise and -6 dB TIR, which relates the noise and co-channel types of interference.

The previous tests with human listeners showed 96 % correct for no interference [17]. This is almost the same as our tests with automated speech recognition, which indicates that the SPHINX [3] software is operating as well as humans using clean speech.

The previous human TIR tests yielded a 54 % success rate at -12 dB TIR [6]. Our test yielded about the same success rate (50 %) at -6 dB TIR, or an 18 dB change from the human tests. The previous human TIR tests yielded a 73 % success rate at -6 dB TIR, while our test had a similar success rate (74 %) at 12 dB TIR, or again, an 18 dB change from the human listing test. This delta differs quite a bit from the previous finding of a 23 dB difference from man to machine. This could be the result of a better machine being used now than was available at the time of the previous comparison.

Comparing the noise tests to the co-channel tests, at a given interference level, shows that noise causes greater difficulty in

recognition success. This result is similar to work done previously that used automated speaker identification [4]. Analyzing the data, it was found that at very high noise interference levels, the output of the SPHINX software indicated that no word had existed. This was not the case with the co-channel interference. This could be due to the following. With co-channel interference there is a silent period at the start of each utterance, which might trigger the SPHINX software into assuming a word is about to start. In the method used to add noise, no such period of silence existed. There is also a silent period at the end of co-channel speech that the SPHINX software might use to signify the end of a word. Again there was not such silent period at the end of the noise mask utterances. With even a random guess at some word, co-channel results should be improved over not taking any guess. It is also noted, that the differences in percent correct between the noise and co-channel cases increase with increasing interference levels. This could be due to differences in the actual interfering levels used (due to differing calculation methods) or a different effect on the SPHINX software caused by the two types of interference. Using the MNRU (word masking method) may have resulted in less discrepancy between using noise or co-channel masking. Because further work will be using the co-channel data, the noise case is not being pursued.

Another interesting observation is that the SPHINX software can handle corrupted speech (either by noise or co-channel interference) better with men targets than women targets, without regard to the sex of the interferer. For instance, at 0 dB SNR the men had a success rate of 14 %, versus the woman at 6 %, and at 0 dB TIR the men had a success rate of 34 %, versus the woman at 25 %.

The statistics of the error rate was analyzed to see how close the percent errors are being estimated. Using a binomial distribution estimate, with a 95% confidence level, the error rates for the cases of all men or all women or all speakers are no greater than a few percentage points. The test using training data used a much smaller data base than the one above, and the accuracy of the results are not as reliable.

During the next phase of this project we will try various reconstruction methods to fill in the unusable speech segments. These segments will be extracted from the co-channel speech. The research will be done with the SPHINX software as an analysis tool, using this method to compare reconstruction methods to each other and to the base lines established here.

A future evaluation would be to use the same database and error definition with human listeners. This could then be used to find a possible constant correction factor to transfer results of automated testing to human testing.

REFERENCES

- [1] ANSI S3.2 "Method for Measuring the Intelligibility of Speech over Communications System," **American National Standards Institute**, 1999.
- [2] Nishant Chandra and Robert E. Yantorno, "Usable Speech Detection Using the Modified Spectral Autocorrelation Peak To Valley Ratio Using the LPC Residual", **4th IASTED International Conference Signal and Image Processing**, 2002, pp:146-150.
- [3] School of Computer Science, Carnegie Mellon University, SPHINX-II, 2001. [Online], Available at <http://fife.speech.cs.cmu.edu/sphinx/>
- [4] Sharon Gannot, David Burshtein, and Ehud Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," **IEEE Transactions on Speech and Audio Processing**, Volume: 6 Issue: 4, July 1998, pp. 373-385.
- [5] John S. Garofolo, et. al., "DARPA TIMIT: acoustic-phonetic continuous speech corps CD-ROM," **U.S. Department of Commerce**, 1993.
- [6] Brian A. Hanson and David Y. Wong, "Processing techniques for intelligibility to speech with co-channel interference," **Signal Technology, Inc.**, Goleta, CA., Final Technical Report, RADC-TR-83-225, 1983.
- [7] ITU-T, telecommunication standardization sector, P-810, "Modulated Noise Reference Unit (MNRU)," 1996.
- [8] Arvind R. Kizhanatham, Robert E. Yantorno, and Brett Y. Smolenski, "Peak Difference of Autocorrelation of Wavelet Transform (PDAWT) Algorithm Based Usable Speech Measure", **SCI**, 2003.
- [9] Kasturi R. Krishnamachari, Robert E. Yantorno, Daniel S. Benincasa, and Stanley J. Wenndt, "Spectral Autocorrelation Ratio as a Usability Measure of Speech Segments Under Co-channel Conditions", **ISPACS**, 2000, pp:710-713.
- [10] R. Gary Leonard, "A database for speaker-independent digit recognition," **Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing**, 1984, pp. 42.11.1-42.11.4.
- [11] Jereme M. Lovekin, Kasturi R. Krishnamachari, Robert E. Yantorno, Daniel S. Benincasa, and Stanley J. Wenndt, "Adjacent Pitch Period Comparison (APPC) as a Usability Measure of Speech Segments Under Co-channel Conditions", **IEEE International Symposium on Intelligent Signal Processing and Communication Systems**, November 2001, pp:139-142.
- [12] Jereme M. Lovekin, Robert E. Yantorno, Kasturi R. Krishnamachari, Daniel S. Benincasa, and Stanley J. Wenndt, "Developing Usable Speech Criteria for Speaker Identification Technology", **Proceedings of ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing**, 2001, pp:421-424.
- [13] G. F. Meyer, F. Plante, and F. Bethommier, "Segregation of concurrent speech with the reassignment spectrum," **Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing**, 1997, pp. 1203-1206.
- [14] J. A. Naylor and S. F. Boll, "Techniques for Suppression of an Interfering Talker in Co-channel Speech," **Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing**, 1987, pp. 205-208.
- [15] Brett Y. Smolenski and Robert E. Yantorno, "Fusion of usable speech measures using quadratic discriminate analysis", (in review), 2003.
- [16] Brett Y. Smolenski and Robert E. Yantorno, "Fusion of Co-channel speech measures using independent components & nonlinear estimation", **ISPACS**, 2002.
- [17] William D. Voiers, "Uses of the Diagnostic Rhyme Test (English Version) for Evaluating Multilingual Operability in Aviation Communications: An Exploratory Investigation," **Multi-lingual Interoperability in Speech Technology**, 1999.
- [18] Robert E. Yantorno, K. R. Krishnamachari, Jereme M. Lovekin, D. S. Benincasa, and S. J. Wenndt, "The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Employed as a Co-channel Detection System", **IEEE Workshop on Intelligent Signal Processing**, Hungary, May 2001, pp: 193-197.

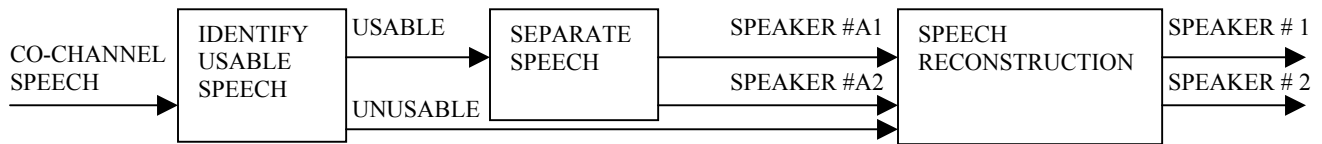


Figure 1: Block Diagram of a Co-channel Speech Reconstruction System

Table 1: Summary of other work. Percent correct vs. SNR/TIR

		inf. dB	5 dB	0 dB	-6 dB	-12 dB	-18 dB
Human, TIR				88	73	54	
Human, MNRU		96		72			
Threshold, TIR			machine				human

Table 2: Percent correct for each speaker vs. SNR

		inf.dB	18dB	12dB	6dB	0dB	-6dB
Men		98	88	73	39	14	1
Women		96	78	53	25	6	0
Average		97	83	63	32	10	0

Table 3: Percent correct for trained speaker vs. SNR

Noise type		30 dB	24 dB	18 dB	12 dB	6 dB	0 dB
Level	Men	69	61	61	57	41	26
	Women	30	17	5	5	8	1
	Average	50	39	33	31	24	14
Modulated	Men	84	85	80	37	22	0
	Women	72	65	58	48	30	15
	Average	78	75	69	42	26	8

Table 4: Percent correct for each category vs. TIR

Target	Interferer	inf.dB	18dB	12dB	6dB	0dB	-6dB
Men	men	98	93	79	59	38	22
Men	women	98	91	79	57	30	19
Average men		98	92	79	58	34	21
Women	men	96	82	69	44	27	18
Women	women	96	85	68	40	24	16
Average women		96	84	68	42	25	17
Average		97	88	74	50	30	19