

EXPERIMENTAL FACILITY FOR MEASURING THE IMPACT OF ENVIRONMENTAL NOISE AND SPEAKER VARIATION ON SPEECH-TO-SPEECH TRANSLATION DEVICES

Douglas A. Jones¹, Arvind Jairam¹, Wade Shen¹, Paul Gatewood², John Tardelli² and Michael Emonts³

¹MIT Lincoln Laboratory

²ARCON Corporation

³Defense Language Institute Foreign Language Center

ABSTRACT*

We describe the construction and use of a laboratory facility for testing the performance of speech-to-speech translation devices. Approximately 1500 English phrases from various military domains were recorded as spoken by each of 30 male and 12 female English speakers with variation in speaker accent, for a total of approximately 60,000 phrases available for experimentation. We describe an initial experiment using the facility which shows the impact of environmental noise and speaker variability on phrase recognition accuracy for two commercially available one-way speech-to-speech translation devices configured for English-to-Arabic.

Index Terms— Machine Translation for Speech

1. FACILITY FOR LABORATORY EXPERIMENTS

In this report, we introduce a laboratory facility designed to test speech-to-speech (S2S) machine translation devices. We have constructed a test corpus of human speech recordings to be used for S2S evaluation, with special emphasis on military needs. In this facility we are able to conduct experiments in a highly controlled fashion. The speech samples are played out over a calibrated artificial Head And Torso System (HATS) with the speech translation device held in controlled configurations, either at a fixed distance from the sound source or in conjunction with a close-talking microphone. Highly distracting noise can be played over speakers in a sound-proof room at noise levels matched to the speech samples to test the impact of noise on device performance.

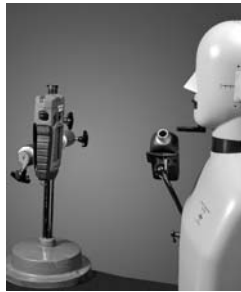


Figure 1: S2S Testing

2. SPOKEN MILITARY PHRASE CORPUS

This corpus is designed to be used with translation systems that take spoken English words and sentences as input and produced prerecorded spoken foreign words and phrases as output. Variation in speaker pronunciation [1] and noise in the environment [2] are two common sources of error for such devices, causing them to fail to translate or to produce incorrect translations. Previously available speech libraries did not meet the evaluation needs for currently available speech translation devices, since most of these devices only recognize a limited set of input sentences and phrases.

2.1. Selection of 1500 English Phrases

We constructed a set of phrases that can be used both for testing current one-way S2S technology and for testing more advanced capabilities. One-way devices accept input in one language, e.g., English, and produce output in another language, e.g., Arabic, but not in the reverse direction.

The phrase list needed to be such that we could record a large quantity of phrases in a half-day recording session for each individual speaker. By maximizing throughput in the recording session, we were able to design a recording protocol capable of capturing approximately 1500 phrases per speaker. Some speakers required multiple recordings for a given phrase, which slowed throughput in these cases. If a given speaker was able to finish the list, we recorded duplicates spoken under additional noise conditions for that speaker.

The phrases were drawn from an intersection of phrases available on two commercial S2S devices and phrases from the Defense Language Institute FLC's Survival Kits [3]. The corpus includes general purpose phrases ("Good Morning."; "Thank you for talking with me.") as well as more specialized phrases in military topics for communication control, search and identification, etc. ("Are members of your immediate family here?", "Has any one in your unit had individual military training?", "Please park your vehicle over there.", etc.)

* This work is sponsored by the Defense Language Institute Foreign Language Center (FLC) under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2.2. Studio Recordings for 42 English Speakers

We recruited subjects to represent a variety of accented and unaccented American English speech, with representation of both male and female speakers. We recorded the self-reported identifications of the subjects who responded to general advertisements in the Boston area for participation in our half-day recording sessions: Males: 4 Asian, 6 Hispanic, 2 African American, 18 white/other; Females: 4 Asian, 2 Hispanic, 2 African American, 6 white/other. Subjects were asked to read the phrases naturally and did not always exhibit salient accents.

In order to evaluate the utility of one-way S2S systems, which encode fixed lists of phrases, it was necessary to collect speech utterances suitable for the two devices in our preliminary experiment.

The recordings for our experiment were tailored to the existing speech translation devices and were produced in three types of acoustic noise environments using a technique that allows for realistic simulation of multiple noise fields that take into account Lombard effects.

During recording, speakers were seated in an acoustic isolation room and were asked to read phrases while wearing a headset with (1) a boom-mounted high quality microphone set at a fixed distance from the lips and (2) calibrated circumaural headphones, with either (a) no noise, (b) low pink noise (65 dB) or (c) high pink noise (75 dB) playback in their headphones. A sidetone was also provided in the headphones at a level 26dB below that presented at the microphone. This value was arrived at by playing a reference tone through the HATS mouth and noting the difference in SPL (Sound Pressure Level) between a position directly in front of the HATS mouthpiece and a simultaneous measure using the HATS ear. The consistent sidetone level allowed the speakers to gauge and adjust their vocal effort across the 3 noise conditions to be able to hear themselves adequately. The circumaural headphone design was used to control leakage of the noise playback or sidetone through the earcups and into the microphone.

The noise presentations elicited stressed speech patterns resulting from Lombard effects (i.e. Low Lombard, High Lombard, and No Lombard). The data was recorded in a sound-proof studio so as to minimize ambient noise and room effects, and phrases were recorded in each Lombard condition. Each recording session was marked with a 95 dB reference tone for playback calibration.

This process allows clean speech data (with Lombard effects) to be mixed at test time with recorded noise field data to simulate Lombard-affected noisy speech for presentation to each device during testing. This protocol also allows us to tease apart Lombard effects and additive noise effects.

2.3. Sound-proof Laboratory for Device Testing

To conduct experiments, the recorded speech was played to the S2S devices through a HATS in a sound-proof studio. The HATS meets ITU P.58 specifications and is calibrated accordingly [4]. Playback calibration was performed using a reference tone generated per recording session per speaker. The ability of the device to recognize each phrase was then measured. In both low and high noise conditions, speech babble noise from the NoiseX corpus was played over speakers in the sound studio at noise levels matched to those presented to subjects during recording [5]. We measured the peak signal-to-noise ratio of the intermittent noise for the low noise and high noise conditions at 14 and 8 dB respectively.

3. INITIAL EXPERIMENTS

As a preliminary experiment, we measured the impact of environmental noise and speaker gender on phrase recognition accuracy for two commercially-available one-way speech-to-speech translation devices. Because the primary purpose of this report is to demonstrate the laboratory facility and the methods for assessing devices, rather than to compare these two specific devices, we will refer to them simply as Device A and Device B.

In this preliminary experiment, we used a small dataset in order to observe which effects are robustly measurable with a minimal amount of laboratory experimentation time. This experiment used 27 phrases spoken by 42 English speakers in three noise conditions: quiet, low noise (65dB) and high noise (75dB). We have conducted a preliminary experiment using a selection of phrases from the topic set for Identification of Family History, which includes phrases such as “Are members of your immediate family here?”, “Where did you last see them?”, “Do you know where your family is?”, etc. Even with this small dataset, we were able to observe robust results for the impact of noise and speaker gender. Our expectation is that the results we report here can be replicated elsewhere, using a similar experimental facility.

3.1. Environmental Noise

In this experiment, three noise conditions were tested: Quiet, Low Noise and High Noise. For the noise conditions, speech babble noise was played over speakers in the sound studio at noise levels matched to those presented to subjects during recording. Both devices were held at a fixed distance of 12 inches from the HATS mouth. For Device A, in quiet conditions, 93% phrase recognition accuracy was obtained. In low and high noise conditions, the device was accurate 90% and 78% of the time respectively, as shown in Figure 2. For Device B, the accuracy was 93%, 80% and 50% for these conditions

respectively. The error bars indicate standard error for the data samples in each condition. Phrases had to match exactly in order to be scored as correctly recognized.

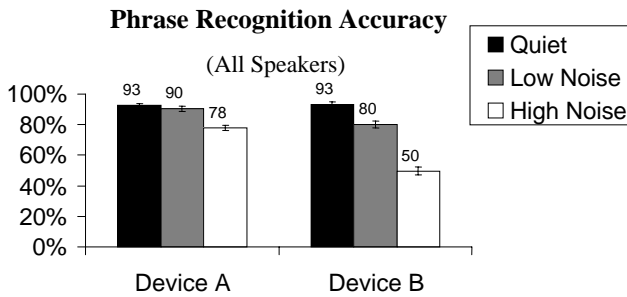


Figure 2: Impact of Noise on Phrase Recognition

The reason for holding the S2S devices at a fixed distance of 12 inches is that this is a comfortable distance that allows a user to see the hand-held screen easily. However, for noisy environments, both device manufacturers recommended either using a close-talking microphone or holding the device closer to the mouth. We conducted an additional experiment using properly adjusted close-talking microphones for each of the devices, with the microphones situated 0.5 inches from the lips of the HATS and off to one side, and we observed results only slightly worse in the high noise condition as compared with the quiet condition, as shown in Figure 3.

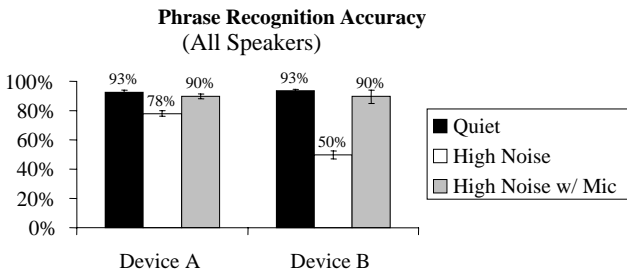


Figure 3: Compensating for Noise with Microphone

3.2. Speaker Variation

Figure 4 shows the overall speaker variation in terms of phrase recognition accuracy, averaged over all noise conditions for both devices. Although there is substantial variability for the speakers, there are systematically worse recognition results for female speakers, shown at the right end of the graph in the darker bars, compared with the male speakers at the left end in lighter bars. The error bars indicate the standard error.

In fact, the distinction between male and female speech is a very robust effect. Figure 5 shows the impact of speaker gender on phrase recognition accuracy. For Device A, the phrase recognition accuracy across all noise conditions was 91% for males and 75% for females. We

also see better results in phrase recognition for males with Device B: 78% for males and 64% for females.

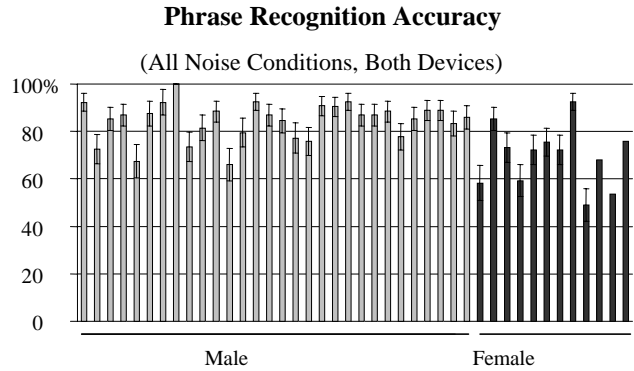


Figure 4: Overall Speaker Variation

Moreover, we see that male speech is recognized better in all noise conditions for both devices. For Device A, the accuracy for males versus females in Quiet, Low Noise and High Noise was 96% vs. 84%, 95% vs. 82%, and 83% vs. 60%, respectively. For Device B we observed 96% vs. 86%, 85% vs. 70%, and 54% vs. 33%, as shown in Figure 5.

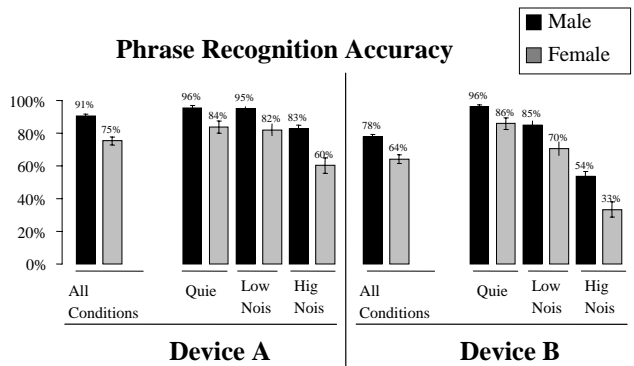


Figure 5: Impact of Gender on Phrase Recognition

The difference between male and female recognition may require advances in S2S system design, depending on how much of this effect is due to training data issues and how much is due to deeper issues of speech coding. We tested whether the difference could be due to females not speaking as loudly as males, but we found this not to be the case. Figure 6 shows the distribution of speakers by recognition rate, peak speech signal level and gender for device B. Speaking louder did not correspond to higher recognition accuracy. Similar results hold for device A. Average loudness for females in quiet, low noise and high noise conditions were 77, 80, and 84 dB respectively, and 77, 79 and 83 dB for males.

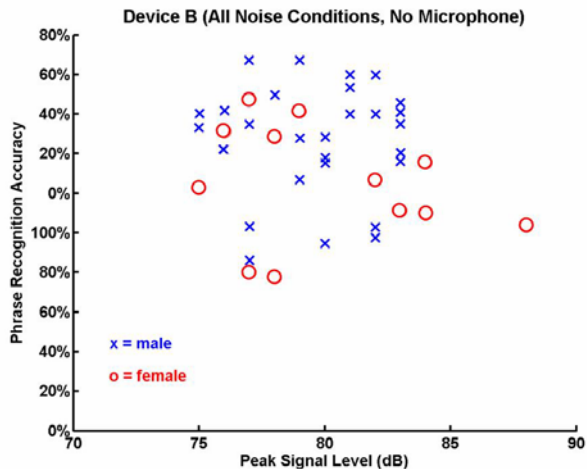


Figure 6: Impact of Loudness on Phrase Recognition

We observed that once the speech volume was calibrated to the proper level, there was no additional benefit for phrase recognition from speakers who spoke in a louder voice than those who spoke less loudly.

In fact, we observed that stressed speech (i.e., speech with the induced Lombard effect) that was spoken over 75 dB noise in the sound studio and played in the presence of 75 dB background noise in the S2S lab received numerically worse results than unstressed speech (i.e., speech spoken over silence in the sound studio) played in the presence of the same 75 dB noise, when a microphone is used in both cases. In future work we will explore the specific impact of the Lombard effect more fully.

5. CONCLUSIONS

In order to assess the recognition accuracy for the diverse regional and ethnic backgrounds of U.S. armed service personnel, we recorded a set of approximately 1,500 phrases. These phrases were spoken by speakers recruited to reflect the demographic speech patterns of the personnel who would be using the device operationally. The speech corpus includes modest variation for speaker gender and ethnic accents of American English, including Spanish-accented English. For a preliminary experiment, we selected a very small sample of phrases from this larger set of phrases to measure the robust effects of environmental noise and speaker gender. We observed that negative impact of environmental noise in our experiments can be mitigated by using a close-talking microphone with the devices. We also observed that the two devices recognized female speech more poorly than male speech.

6. FUTURE WORK

Our preliminary experiment did not focus on the translations in the devices themselves. Our rationale for

focusing on speech recognition was that we expect the one-way translation devices to employ a simple phrase lookup procedure, in which pre-recorded and pre-translated foreign language phrases are associated with a fixed inventory of English phrases. In future work, we will specifically assess the accuracy of the translated phrases. In the experiment, Arabic language experts will assess both the text and the audio to determine the accuracy of the actual recorded phrases.

The specific impact of the Lombard effect on device performance is more subtle than environmental noise and speaker accent. In future work we will use a larger selection of our recorded data to study these and other effects.

Along those lines, we conducted a preliminary analysis of speaker accent. As might be expected from related work [2], we observed a numeric trend showing slightly worse performance for Spanish-accented speech than for unaccented speech, but we require a larger dataset to produce statistically significant results. We will conduct similar analyses for speakers with various accent types.

7. ACKNOWLEDGMENTS

We especially wish to thank Neil Granoien for providing the opportunity and funding for this work. We wish to acknowledge Melissa Holland at the Army Research Laboratory, for sharing her expertise and identifying the need for the type of laboratory facility we have created here; Tim Anderson, from the Air Force Research Laboratory, on our corpus design and experimental methodology; Elaine Marsh, from the Navy Research Laboratory, for helpful comments drawing on her extensive background on research in speech in noisy environments; Ed Cerutti at US Army Intelligence Command for providing visibility into operational testing needs; and our colleagues at Lincoln who have shared their expertise and sound-room facilities, in particular Michael Brandstein and Kevin Brady.

8. REFERENCES

- [1] Pearce, David and Hirsch, Hans-Günter. 2000. "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", In ICSLP-2000, vol.4, 29-32.
- [2] Ikeno, Ayako, Bryan Pellom, et al. 2003. Issues in Recognition of Spanish-Accented Spontaneous English. IEEE/ISCA Workshop on Spontaneous Speech Processing and Recognition, Tokyo.
- [3] Defense Language Institute Language Survival Kits. 2006. Available at <http://www.lingnet.org/>.
- [4] RSG-10 NOISEX-92 database, speech babble in canteen, 100 people, 88 db(A). <http://www.speech.cs.cmu>.
- [5] ITU P.58: Head and torso simulator for telephony, Approved August 1996.