Comparison of Two Voice Synthesis Systems as to Speech Intelligibility

Jeff A. Lancaster, M.S., Gary S. Robinson, Ph.D., and John G. Casali, Ph.D.

Auditory Systems Laboratory
Human Factors Engineering and Ergonomics Center
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University (Virginia Tech)
Purpose: to investigate the intelligibility of two text-to-speech (TTS) engines at various speech-to-noise (S/N) ratios within aircraft engine noise:

1. DECTalk (v4.5) ‘Perfect Paul’: an ‘older’ TTS engine that has demonstrated superior performance in previous intelligibility studies (e.g., Greene, Logan, & Pisoni, 1986).

2. AT&T’s Natural Voices (1.4) ‘Mike’: a ‘newer’ TTS engine that uses concatenated speech synthesis and was developed with modern technologies.
• **Rationale:** utilize the Modified Rhyme Test (MRT, ANSI 3.2-1989) to measure intelligibility:

1. Research exploring ‘newer’ TTS engines (e.g., AT&T’s Natural Voices) with those that have demonstrated superior intelligibility in the past (e.g., DECTalk) is lacking.

2. Voice technologies have been suggested (e.g., Rehmann 1996, 1997) as an attractive alternative to radio within the cockpit (e.g., airborne data link).

3. Comparison of ‘newer’ and ‘older’ TTS engines for possible integration into advanced aviation systems is thus warranted.
• **Apparatus:**

TTS engines

1. DECtalk v4.5 using voice ‘Perfect Paul.’
2. AT&T Natural Voices v1.4 using voice ‘Mike.’

Flight Simulator

1. i-GATE (integrated General Aviation Training Environment) simulator.

**Active Noise-Reduction (ANR) Communications Headset**

1. Bose ANR Aviation Headset Model AHX-02.
Apparatus (cont.):

- Bose ANR Aviation Headset displayed on an artificial head (ANSI 3.2–1989).

- Participants sat facing the Infinity loudspeakers through which the aircraft engine noise was presented.
• **Hypothesis:** AT&T’s Natural Voices v1.4 ‘Mike’ will result in superior intelligibility, at every S/N ratio, over DECTalk v4.5 ‘Perfect Paul’ in aircraft engine noise.

• **Methodology:** Both TTS engines produced MRT stimuli that were analyzed and equalized for level:

1. Both TTS engines exist as part of the Fonix iSpeak (v3.0) speech synthesis suite, which was installed on a desktop PC.
2. Each TTS engine was set to produce utterances at 60% speed (speech rate).
3. Cool Edit Pro (v1.2a) editing software was used in concert with a Larson-Davis 3200 Spectrum Analyzer (broadband, A-weighted) to analyze the utterances.
4. Each utterance was measured, adjusted, and verified for level and thus equality across TTS engines (per Kryter, 1985).
**Procedure:** Ten participants *(not pilots)* were screened via audiogram per ANSI 3.2-1989.

- **Familiarization Session:** participants completed the MRT using recorded human speech at three speech-to-noise ratios:
  - MRT: includes a list of 50 six-word sets as stimuli:
    
    | way     | may    | say    |
    |---------|--------|--------|
    | pay     | day    | gay    |

  - Three speech-to-noise ratios were presented to the participants: -5 dB, -8 dB, and –11 dB.
  - The noise utilized was Cessna 172R engine noise at 85 dB(A).
  - Participants completed familiarization until they were comfortable with the procedure *and* the experimenter was satisfied with their performance.
**Experiment:** 3x2 within-subjects design

- **Independent variables:**

  1. **TTS engine** (DECtalk v4.5 ‘Perfect Paul’ and AT&T Natural Voices v1.4 ‘Mike’): Each synthesizer was tested on different days and presentation order was alternated for each subject.

  2. **Speech-to-noise level:** -5 dB, -8 dB, and –11 dB (twice at each S/N, averaged)

- **Dependent variable:** the percent correct (%) of marked words, adjusted for chance or guessing, per the MRT procedure (ANSI 3.2-1989):

  \[ R_a = R - \frac{W}{n - 1} \]

- Subjective impressions of each TTS engine were also gathered.
• **Results:**

1. Raw scores (percentages, ranging between 0 and 100%) were entered into the MRT formula to produce an MRT score (per ANSI 3.2-1989).

2. Analysis of variance (ANOVA) revealed a significant main effect of TTS engine on intelligibility ($p < 0.0001$), with the AT&T product performing superior to the DECTalk product.

3. ANOVA also revealed a significant main effect of S/N on intelligibility ($p < 0.0001$). Post-hoc analyses (using the LSD test) indicated significant differences in intelligibility ($p < 0.05$) between each S/N ratio.

4. Linear contrast analysis revealed a significant linear trend ($p < 0.0001$), suggesting that the observed differences in intelligibility depended on the S/N ratio.
Main Effect of TTS Engine

Percent (%) Intelligibility

AT&T Natural Voices 'Mike'  DECTalk 'Perfect Paul'

Speech Synthesizer

78.5
59.6

©Jeffrey A. Lancaster, Virginia Tech, 2004
Main Effect of Speech-to-Noise Ratio (S/N)

Speech-to-noise Ratio (S/N) in Decibels (dB)

Percent (%) Intelligibility

- A: 81.95
- B: 74.75
- C: 50.65

©Jeffrey A. Lancaster, Virginia Tech, 2004
• **Discussion**

1. Subjective impressions indicated a preference for the AT&T product:
   - "The AT&T synthesizer sounded much more realistic and human."
   - "The DECTalk synthesizer’s speech was ‘choppy’.”
   - "The AT&T synthesizer was better than the DECTalk one, but it still has issues, such as garbling words and not stressing some consonants."

2. Implementation of airborne data link using speech synthesis:
   - Loss of the traditional ‘party line’ may have workload and/or situation awareness issues that need to be investigated.
   - Realistic (i.e., human) utterance of such information is thus imperative.
   - Synthesized and/or digitized speech output of ATC commands or traffic advisory information is an attractive alternative to radio.
   - More research is needed to investigate candidate auditory displays.
Reference


