Methods for Subjective Testing of Text-to Speech Systems

Presenter: Amen Hussain

21st January, 2014	Center for Language Engineering (CLE)

Methods of Subjective Assessment

- Segmental Evaluation
  - Diagnostic Rhyme Test
  - Modified Rhyme Test
  - Bell-Core Tests
- ESPRIT-S_{AM} Project
- ITU P.85 Recommendation
- Blizzard Challenge

Segmental Evaluation

- Diagnostic Rhyme Test (DRT)
  - A carrier sentence containing single syllabic word (CVC)
  - Modify one feature of initial consonant
  - Give the listener multiple options of the heard word
- Modified Rhyme Test
  - Modify one feature of initial and final consonant
- Bell-core Tests
  - Evaluation of the intelligibility of sequences of one or more consonants in initial and final word position

Feature

- Place of Articulation
  - Bilabial
  - Dental
  - etc
- Manner of Articulation
  - Stop
  - Fricative
  - etc
- Voicing
  - ب پ
- Aspiration
  - بھ پھ

	Bilabial	Libiodental	Dental	Alveolar	Retroflex	Palatal	Velar	Uvular	Glottal
Stop	P P_H B B_H		T_D T_D_H D_D D_D_H		T T_H D D_H		K K_H G G_H	Q	Y
Fricative		F V		S Z_Z		S_H	X G_G		H
Affricate						T_S T_S_H D_Z D_Z_H			
Nasal	M M_H			N N_H			N_G N_G_H		
Lateral				L L_H					
Approximant						J J_H			
Trill				R R_H					
Tap/Flap					R_R R_R_H				

Examples

- DRT
  - ماپ باپ
  - MA_AP BA_AP
  - CVC CVC
- MRT
  - داغ باگ
  - DA_AG_G BA_AG
  - C V C C V C
- Consonant Cluster Identification
  - تحقیقات
  - T_DAHKI_IKA_AT_D T_DAHGI_IKA_AT_D
  - C VCC V C V C C VCC V C V C

ESPRIT-S_{AM} Project

- Standard Segmental Test
  - Single Syllabic word of the structure CV, VC, and VCV
  - Comprising all phonotactically permissible combinations of initial, medial, and final consonants and three point vowels, e.g., /i/, /u/, and /a/
  - The generated words are often meaningless but they can be meaningful
  - Examples: *pa, ap, apa*
- Cluster Identification Test
  - Single Syllabic word containing consonant cluster and vowel cluster e.g.(CCVCC, VCC,CVVC)

  - Words are generated by considering phonotactical rules they are often meaningless but by chance can be meaningful
- Semantically Unpredictable Sentences
  - Comparative evaluation of sentence intelligibility, minimizing the effect of contextual cues
  - Short, semantically unpredictable sentences of five different, common syntactic structures with words randomly selected from lexicons with frequent "mini-syllabic" words (smallest words available in a given category):
  - Subject – Verb – Adverbial, e.g., *The table walked through the blue truth*

  - Fifty sentences (10 per structure) are recommended per synthesizer.

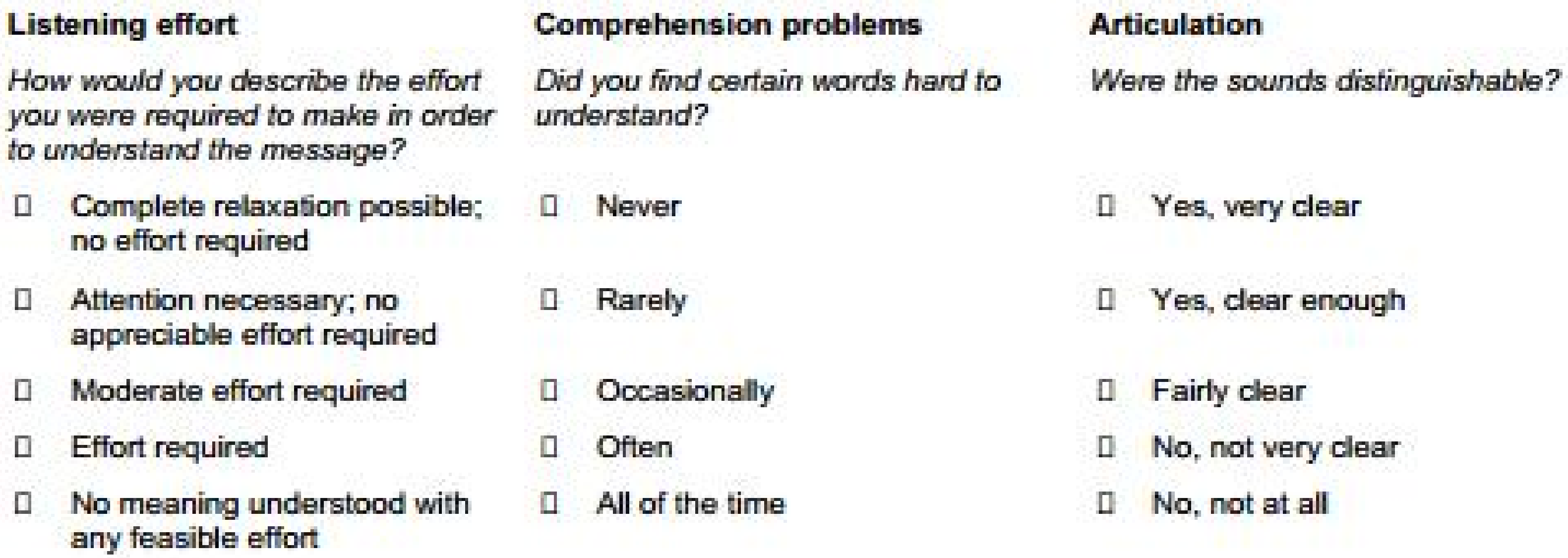
- The overall S_{AM} Quality
  - Comparative evaluation of overall quality aspects, particularly acceptability, intelligibility, and naturalness, for longer stretches of speech.
  - Example: *I realize you're having supply problems, but this is rather excessive and I need to arrive by 10.30 a.m. on Saturday*.
  - Each aspect of speech is rated by a different group of subjects (minimally ten)

ITU P.85 Recommendation

- Multiple Sources
  - Synthesized Speech
  - Degraded Natural Speech
- Speech Material
  - Long Sentences (10-30) seconds
  - Sentences should be from one topic
  - Example: Miss Robert, the running shoes color: white, size: 11, reference: 501-97-52, price: 319 francs, will be delivered to you in 1 week.

- Evaluate Naturalness
  - Pronunciation
  - Speaking Rate
  - Voice Pleasantness
- Evaluate Intelligibility
  - Listening Effort
  - Comprehension Problems
  - Articulation
  - Fill in the blanks from the content heard

Example: Evaluate Intelligibility



Listening effort

How would you describe the effort you were required to make in order to understand the message?

□ Complete relaxation possible; no effort required

□ Attention necessary; no appreciable effort required

□ Moderate effort required

□ Effort required

□ No meaning understood with any feasible effort

Comprehension problems

Did you find certain words hard to understand?

□ Never

□ Rarely

□ Occasionally

□ Often

□ All of the time

Articulation

Were the sounds distinguishable?

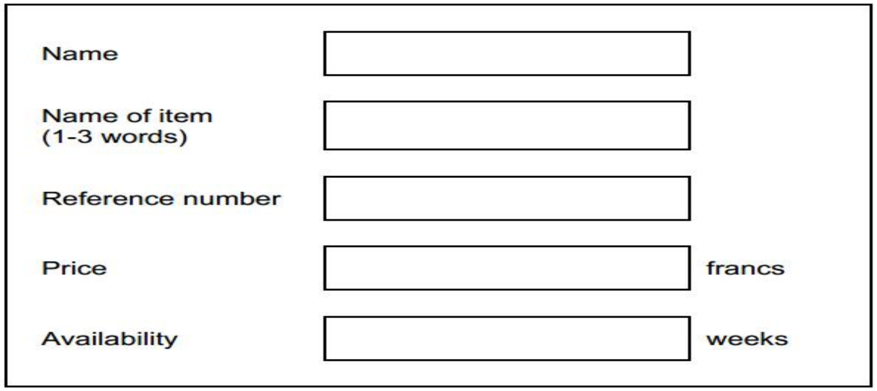
□ Yes, very clear

□ Yes, clear enough

□ Fairly clear

□ No, not very clear

□ No, not at all



Name		
Name of item (1-3 words)		
Reference number		
Price		francs
Availability		weeks

- Rank overall Quality
- Acceptability Test

Blizzard Challenge

- Speech Material
  - From five different genres
    - Novel
    - News
    - Conversations
    - Semantically Unpredictable Sentences (SUS)
    - Phonetically Confusable Sentences (DRT/MRT)

- Naturalness Evaluation
  - MOS (Mean Opinion Score)
    - Rank the overall speech quality on the scale of 1-5 from first three genres
- Intelligibility Evaluation
  - Write the sentences heard from last two genres

Blizzard Challenge Survey

2005	2007	2008	2009	2010	2011	2012
Naturalness News	Naturalness News	Naturalness News	Naturalness News	Naturalness News	Naturalness News	Naturalness News
Naturalness Novel	Multidimensional Scaling	Naturalness Novel	Multidimensional Scaling	Naturalness Novel	Naturalness Novel	Naturalness Novel
Intelligibility SUS (WER)	Intelligibility SUS (WER)	Intelligibility SUS	Intelligibility SUS	Intelligibility SUS (clean)	Intelligibility SUS	Intelligibility SUS (WER)
Intelligibility Phonetically Confusable (DRT/MRT)	Similarity Test	Similarity Test	Similarity Test	Similarity News	Similarity News	Similarity Novel
Naturalness Conversational	Naturalness Conversational		Naturalness Conversational	Intelligibility SUS (noise)	Similarity Novel	Multiple dimensions testing
			MOS Appropriateness		Intelligibility Address	
					Naturalness Reportorial	

Methods for Subjective Testing of Text-to Speech Systems

- Multidimensional Scaling
  - In each part, listeners heard pairs of different sentences – one sample from each of two of the participating systems, or, in the case of one system ordering for each dataset, two samples from the same system.
  - Listeners were to ignore the meanings of the sentences and instead concentrate on how natural or unnatural each one sounded. They then chose whether, in their opinion, the two sentences were similar or different in terms of their overall naturalness.
- MOS Appropriateness
  - Listeners saw a question (provided in text form only) of the type that a human user might ask a restaurant enquiry service, and then listened to one spoken sample that represented the response to that question. Listeners chose a score which represented how appropriate or not the response sounded in that dialogue context on a scale of 1 [Completely Inappropriate] to [Completely Inappropriate]

- Multiple dimensional testing
  - Overall impression ([bad] to [excellent])
  - Pleasantness ([very unpleasant] to [very pleasant])
  - Speech Pause ([speech pauses confusing/unpleasant] to [speech pauses appropriate/pleasant])
  - Stress ([stress unnatural/confusing] to [stress natural])
  - Intonation ([melody did not fit the sentence type] to [melody fitted the sentence type])
  - Emotion ([no expression of emotions] to [authentic expression of emotions])
  - Listening effort ([very exhausting] to [very easy])

More Tests

- Minimal Pair Intelligibility Test
  - Words can differ in one or two features
  - MPI test data contains consonants and vowels, onsets, nuclei and/or codas, consonant clusters, mono-syllabic and poly-syllabic words, and stressed and unstressed syllables
- Phonetically Balanced
  - Phonetically balanced words in a carrier sentence
  - phonetically-balanced words that use specific phonemes at the same frequency as they appear in language.

- Prosody Evaluation
  - PURR method
    - De-lexicalise the speech stimuli to ensure that the listener perceives only the prosody of an utterance.
    - This is done by reducing the speech signal to produce stimuli that convey only intensity, F0 contour and temporal structure.
  - Human-Machine Prosody Comparison