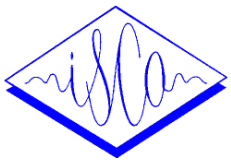

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A FEATURE-PROFILE FOR APPLICATION-SPECIFIC SPEECH SYNTHESIS ASSESSMENT AND EVALUATION

Ute Jekosch, Louis C. W. Pols***

*Lehrstuhl für allgemeine Elektrotechnik und Akustik, Ruhr-Universität Bochum, 44780 Bochum, Germany
**Institute of Phonetic Siences / IFOTT, University of Amsterdam, Herengracht 338, 1016 CG Amsterdam, The Netherlands

ABSTRACT

This paper is concerned with assessing the quality of speech synthesizers in application utilizing the conclusions made by subjects. Ultimately, the quality of speech becomes apparent in its usage, in this case the interaction between a machine and a biological system in a specific communicative context. As it is not possible to attain precise control over environmental circumstances of system application, however, a vast range of complicated human responses is possible. The approach introduced here is an attempt to bridge the gap between theory and experiment on the one hand, and, the actual performance of the system in application on the other hand, while still advocating the opinion that laboratory analogues of real applications are conceivable. What is still missing at present is a meaningful correlation between the properties of the speech signal and their relative relevance in an application-specific pragmatic context. Such a functionalized view of speech requires a unique formal description of both the speech signal properties and independently defined rudimentary variables characteristic of the situational communicative context.

I INTRODUCTION

The aim of the approach described here is to identify a set of clear criteria in order to assess speech output systems in applied contexts. Imagine you are given the task of assessing the quality of a speech output device used as a traffic information system implemented in a noisy car moving with different speed. Neither specific tests nor a set of tests are available for immediate use. However, in such a situation speech assessment experts still know what to do: They analyse the conditions, identify the most important factors influencing speech quality and then they decide which of the available tests is/are appropriate - or which ones have to be modified - for assessing/evaluating these quality aspects with respect to general system performance. In other words, both the application situation as well as the available tests are analyzed with the view of accomplishing the given task. A decision is taken based on this feature analysis to define a test setup. In general, such decisions are not arbitrary but are also supported by other experts.

If in speech assessment experts have solved a specific problem in such a way, it is often extremely difficult for them to justify and explain the reasons for their approach. One reason is that the field of speech assessment has many facets so that there is no clear structure and no straight-forward overview of major components. Based on a structured analysis of speech output tests and their concomitant influence factors, the concept of speech assessment feature profiles will be discussed followed by an example from speech synthesis in application.

II FACTORS OF INFLUENCE

There are a number of factors that might have a specific influence on test results or a general influence on system performance: features regarding synthesizers, the structure of test methods, signal properties, criteria for actual test runs, selection of subjects, and the intention of the assessment expert. In the following sections a broad description of each of these will be given. Without doubt these lists are not complete; they should be understood as a collection of features that can or has to be extended whenever neccessary.

2.1 Synthesizer

- mono-lingual or multi-lingual device
  - which language(s)
- module for input language identification
- concept-to-text, keywords-to-text, input of a scanner, text directly
- restricted vs. open vocabulary
- automatic identification or possibility of manual labeling of foreign language terms in the input text
- automatic identification/possibility of manual labeling of others
- text pre-processing (abbreviations, numbers, tables...)
- grapheme-to-phoneme conversion (rule-based, lexicon-based, rule compiler, accentuation ...)
- linguistic processing
  - morphological analysis (pattern-based, morpheme lexicon, word form lexicon, exceptions lexicon, training lexicon ...)
  - syntactic analysis
  - phrasing (phonological/prosodic structure ...)

SUS:

- available for Dutch, English, French, German, Ialian
- segmental evaluation:
  - test on sentence level
  - test vocabulary:
    - sentence with fixed structure, words to fill in the structure: open C_nVC_n-structure
    - mostly mono-syllabic, except. bi-syllabic words
    - semantic-bearing words
    - phonetically/phonemically balanced (open)
    - frequency of occurrence of test stimuli (sentences) (yes)
    - total number of test stimuli: 50 sentences, of approx. 7 words each, open
    - structure of the test stimuli:
      - words embedded in a sentence
      - form and structure of the embedding unit
        - sentence:
          - declarative, imperative, ...
          - semantically unpredictable
      - frequency of occurrence of target test items: of words (open)
  - any other features: sentences can be automatically generated, vocabulary for generation can be defined test-individually, available for different languages
- measurement based on:
  - transcription, graphemic form: open response
- test form:
  - open response

So much for examples of the test methods themselves. For each particular actual test run, the synthesized signal that reaches the listener's ear can be described according to its specific properties.

2.4 Speech Signal Properties

- pure primary signal (i.e., the synthetic signal)
- primary signal with slight modification
- secondary signal (e.g., noise) where the
  - input-signal modified by
    - channel-characteristics (high quality, telephone - handset, mobile, bandwidth)
    - disruptive factors:
      - stationary signals
      - quasi-stationary signals
      - non-stationary signals (e.g., competing speech, reverberation)
- recording modalities
  - direct output of the synthesizer
  - in an acoustically 'clear' environment
  - in adverse conditions
  - recording equipment
    - microphone
    - analog recorder
    - digital recorder
- presentation equipment
  - analog recorder
  - digital recorder
  - computer, synthesizer, telephone hand-set
- additional presentation of noise for the listener
  - via headphones
  - via loudspeaker
  - in a real environment (office, car, station...)

2.5 Actual Test Run

- test location (sound-proof chamber, office room, free field, ...)
- presentation of speech signals via
  - headphones
  - telephone hand-set
  - loudspeakers, ...
- environmental noise
  - different from the test controlled noise-environment
  - noise-free environment
- stimulus interval
  - fixed
  - variable
- stimulus repetition
  - transmitted n times
  - repeated at subject's request
- stimulus presentation order
  - randomized
  - once defined, per subject same order
- SPL
  - non-constant vs. constant average value
  - determined by the test manager
  - chosen by the subject
  - once chosen, kept constant in test course
  - varied in test course

2.6 Selection of Subjects

- total number
- sex
- age
- education
- normal hearing/hearing impaired
- special 'ear-training' (phonetic training, ...)
- native language
- other languages
- dialect/region where he/she grew up
- dialect/region where he/she lives
- state of cooperation, motivation
- others

2.7 Intention of the Assessment Expert

An analysis of conditions also includes the motivation of the person who wants to run the test. He might be interested in:

- test theory and test application:
  - test enhancement: basic research into test methods and methodologies (test theory)
  - test adequacy: application of test methods and methodologies to check their validity and reliability (assessment of the test behaviour in application)

- synthesizing technique: concatenative technique, parametric synthesis-by-rule, articulatory synthesis)
- synthesis unit inventory: type(s) of units, size
- accentuation and phrasing (phrase and sentence contour)
- processing mode (spelling, single words, paragraphs, ...)
- variety of voices: male/female/child voice, voice adaptation to an individual speaker
- speaking rate, pitch level
- speaking styles (casual, clear, formal, emotional speech)
- speech output combined with other modalities (audio-visual, ...)
- additional system-specific technical details:
  - size, weight, price, interface, modularity, user options

2.2 General Structural Description of Test Methods

- Segmental evaluation:
  - single word tests (SAM-Segmental Tests, Diagnostic Rhyme Test, Cluster-Identification-Test, ...)
  - tests on sentence level
    - Harvard psychoacoustic sentences
    - Haskins semantically anomalous sentences
    - SUS (semantically unpredictable sentences) ...
  - test vocabulary:
    - vocabulary automatically generated or fixed
    - mono-syllabic, bi-, multi-syllabic words
    - semantic-bearing, non-semantic-bearing, semantically unpredicatable
    - vocabulary phonetically/phonemically balanced
    - frequency of occurrence of test stimuli (e.g., words)
    - total number of test stimuli
    - structure of the test stimuli:
      - target item embedded in a larger unit (e.g., initial phoneme in a mono-syllabic word, a single consonant in a CV-structured mono-syllable, a word in a sentence, ...)
      - form and structure of the embedding unit
        - sentence:
          - declarative, imperative, ...
          - semantic-bearing, non-semantic-bearing, semantically unpredictable
      - frequency of occurrence of target test items
      - any other features
- tests on paragraph level
  - comprehension tests
- tests on prosody
  - SOAP protocols for prosodic testing
  - SAM prosodic form test
  - SAM prosodic function test. ...
- overall quality tests
- linguistic and psychological evaluation
  - word verification task, lexical decision task
  - sentence verification task
  - naming task, word recall (serial order)
  - target detection task
  - priming task

2.3 Response Modalities

Measurement based on

- identification/transcription (graphemic or phonemic)
  - open response
  - closed response
- opinions
- score (e.g., Magnitude Estimation, Categorical Estimation)
- choice
  - multiple-choice
  - binary choice (Pair Comparison)
- monitoring the reaction/use
- interview [cf. Delogou et. al]

These feature sets can be used in order to describe available speech assessment tests, e.g., segmental tests:

Modified Rhyme Test (here for the German language):

- available for different languages
- structural description
  - fixed vocabulary
  - single word test
  - mono-syllabic words
  - semantic-bearing
  - phonemically balanced
  - frequency of occurrence of test stimuli
  - number of test stimuli: 900
  - structure of test stimuli: CVC (C being the target entity)
  - frequency of occurrence of target test item (no)
- measurement based on:
  - identification: closed response
- interview after the test (open)

Cluster-Identification-Test:

- available for German
- structural description
  - vocabulary open, automatically generated according to statistical information on mono-syllabic words to be used as test material
  - single word test
  - mono-syllabic words
  - mostly non-semantic-bearing
  - phonemically balanced (open)
  - frequency of occurrence of test stimuli (open)
  - number of test stimuli (open)
  - structure of test stimuli: C_nVC_n, C_nV, VC_n (C_n being the target entity)
  - frequency of occurrence of target test item (open)
- measure based on:
  - identification: open response
- interview after the test (open)

- speech output (SO) systems theory, development and enhancement:
  - SO systems theory and development: Support individual data for basic research into speech output systems enhancement and development (information supplier for SO-experts: 'In what areas do they have to improve their systems?')
  - SO systems adequacy: Provide data for SO system application: Customer: 'Is system x the best choice for my application?'

There are a number of other quite important aspects that could be analysed with respect to their main features (use of speech databases, of reference systems, synthesizers in a dialogue situation, combined with other modalities such as visual support); however, we think that the remaining space should be used to discuss an example.

III SYSTEM TESTING IN APPLICATION

The next step in our approach is to check to what extent these test features can actually be used for synthesis assessment in application. Imagine you have the task of assessing synthesis performance in the context of having to read aloud proper names over the telephone line for any information service in order to tell the SO-system designer how his/her system behaves in a given scenario and where he/she can improve the system's performance. The first step is to analyse the application conditions:

- synthesizer:
  - application: pronunciation of proper names
  - unspecified; test should be run for each device
- structural description of test object (proper names):
  - can be from different language families
  - can be from common semantics (Fisher) or non-semantic bearing
  - frequency of occurrence related to
    - the distribution of names of the different language families within the community
    - the population
    - average everyday use (e.g., 'Clinton' frequently used in the media at present)
    - letter transitions
    - proper name constituents (e.g., syllables, clusters, phonemes)
    - can be known to one subject but not to another

When the feature profile of the application scenario is available the person who wants to run the test can define the best test for this specific application, e.g.,

- selection of tests: at least segmental intelligibility & acceptance test
  - segmental intelligibility: vocabulary should mirror characteristics of the object to be tested
    - response modality: measure based on graphemic transcription, open response
  - acceptance test: best in a task-specific environment (e.g., dialogue, where subject has to welcome fictitious Mr./Mrs. X bearing the announced name)
    - response modality: measure based on score, Categorical Estimation

After this has been done, a comparison of what is desired and what is available (standard assessment tests) can be made. The decision as to whether the best - although not the optimal - fitting standard test will be applied or whether a new test has to be designed (following general standard test methodological considerations) can only be made after weighing up the additional effort necessary to achieve the optimal goal and the sub-optimal solution of simplifying the test object.

Whatever decision is made - the profile for speech quality testing is a helpful guideline in the test construction and preparation phase, and it is also a helpful tool for test documentation [cf. Hegehofer]. The test results can only be meaningful when the tests are performed in comprehensive well-documented conditions.

IV SUMMARY

We have presented a detailed feature profile for speech assessment testing. Without doubt the form and content of this profile have to be refined and gradually improved. Nevertheless, this paper was intended to figure out the basic idea of such a feature profile: To have a tool available for constructing theoretic tests, preparing practical tests and documenting actual test runs. However, standardising such a profile requires much more expertise and effort than we have so far been able to contribute to this approach.

ACKNOWLEDGEMENT

As members of the ESPRIT Projects SAM, SAM_A and EAGLES we would like to thank all colleagues for fruitful discussions which supported and furthered our ideas.

REFERENCES

Benoit, C., Erp, A. van, Grice, M., Hazan, V. & Jekosch, U. (1989), 'Multilingual synthesiser assessment using semantically unpredictable sentences', Proc. Eurospeech'89, Paris, Vol. 2, pp. 633-636.

CCITT (1987), 'Subjective quality assessment of synthetic speech..' CCITT-Contribution XII-176-E

Goldstein, M, Lindström, B, Delogu, Chr., Falcone, M., Sementina, C. (1993), 'Report on structured analysis of speech output assessment tests.' internal paper ESPRIT-Project 6819, SAM_A: Speech Technology Assessment in Multilingual Applications

Hegehofer, Th. (1994), 'A Description Model for Speech Assessment with Subjects.' this issue

Jekosch, U. (1994), 'Speech Intelligibility Testing: On the Interpretation of Results.' Journal of the American Voice I/O Society, 15th vol., p. 63-80

Pols, L.C.W. & SAM partners (1992), 'Multi-lingual synthesis evaluation methods.' in Proc. ICSLP '92, Banff, Vol. 1, p. 181-184