speech synthesis evaluation

nick campbell
atr media information science labs
kyoto, japan

nick@atr.jp

current state of the art

The European taskforce creating human-machine interfaces SIMILAR to human-human communication

SIG on **multimodalities** (WP9)

- State-of-the-art speech synthesis (using automatic selection of units taken from large speech corpora) has very recently reached a level of quality which could not even be foreseen by experts five years ago, thanks to the emergence of speech synthesis techniques based on large speech corpora. The TTS (text-to-speech) industry worldwide has jumped into the new markets opened by this technology, including some European companies and R&D centres, which are now in a position to deliver high-quality speech synthesis especially in term of naturalness. However, despite this global quality improvement, critical acoustic artefacts remain audible in synthetic speech and so there is a real need to develop more sophisticated processing so as to avoid an inhomogeneous output speech quality, both in terms of prosody and spectral continuity. Moreover another strong additional challenge must be addressed: EU synthesis needs to be multilingual. Efforts in these directions therefore need to be reinforced by stronger collaboration of EU partners.

- The real challenge now in TTS technology is that of voice quality control, which includes emotional speech synthesis, as well as voice conversion and adaptation (being able to adapt a high quality synthetic voice to a user's voice or to provide a large panel of differentiable synthetic voices from a single one). All of these topics are commercially motivated by the fact that high quality synthetic speech is currently only reachable through the design of large databases, which take time and money

(from http://www.similar.cc)

frameworks for synthesis evaluation

- **what we CAN do**
  - make machines talk intelligibly
  - with recognisable voices (and faces)
  - within limited computing resources
- **and what we CAN'T do**
  - reproduce conversational speech
  - use expressive non-speech sounds
  - in all the world's languages (and cultures)

the need for continuing incremental evaluations

- for the developer
  - to know the strengths and weaknesses
  - to know which areas need most urgent work
- for the customer
  - to determine the needs of the society
  - to meet their various expectations
- for the science
  - to understand human speech communication

the next paradigm shift . . .

- to meet the expectations of the market:
  - researchers shouldn't define the limits
  - top-down design needs bottom-up input
- we should evaluate systems as much by what they **can't do** as by what they can
  - but this needs imaginative expectations!

Survey of the State of the Art in Human Language Technology (1996)

Editorial Board:
Ronald A. Cole, Editor in Chief
Joseph Mariani
Hans Uszkoreit
Annie Zaenen
Victor Zue

Managing Editors:
Giovanni Battista Varile
Antonio Zampolli (may he rest in peace)

Sponsors:
National Science Foundation
European Commission

13.7 Speech Synthesis Evaluation

Louis C. W. Pols

University of Amsterdam, The Netherlands

- The possibility to generate any existing text, any to-be-worked-out concept, or any piece of database information as intelligible and natural sounding (synthetic) speech is an important component in many speech technology applications [Sor94]. System developers, product buyers, and end users are all interested in having appropriate scores to specify system performance in absolute (e.g., percentage correct phoneme or word intelligibility scores) and in relative terms (e.g., this module sounds more natural for that specific application in that language than another module) [Jek93].

- Since synthetic speech is generally derived from text input (see also chapter 5), not just a properly functioning acoustic generator is required, but also proper text interpretation and preprocessing, grapheme-to-phoneme conversion, phrasing and stress assignment, as well as prosody, and speaker and style characteristics have to be adequate. On all these, and several other, levels one might like to be able to specify the performance, unless one really only wants to know whether a specific task can properly be performed in a given amount of time. This opposes the approach of modular diagnostic evaluation to the one in which global overall performance is the main aim.

(from http://cslu.cse.ogi.edu/HLTsurvey/ch13node9.html)

13.7 Speech Synthesis Evaluation

13.7.1 Modular Diagnostic Evaluation

- At this diagnostic level a suite of tests is already available, although there is little standardization so far, nor are there proper benchmarks. Also comparability of test design and interpretability of results over languages, is a major point of concern [LGP89,Pol91]. The type of tests we have in mind here are methods to evaluate system performance at the level of text pre-processing, grapheme-to-phoneme conversion, phrasing, accentuation (focus), phoneme intelligibility, word and (proper) name intelligibility ([Spi93], performance with ambiguous sentences, comprehension tests, and psycho-linguistic tests such as lexical decision and word recall. There is a great lack of proper tests concerning prosody, and speaker, style and emotion characteristics, but this is partly so because rule-synthesizers themselves are not yet very advanced concerning these aspects either [Pol94b]. However, concatenative synthesis with units taken from large databases plus imitation of prosodic characteristics, is one way to overcome this problem of insufficient knowledge concerning detailed rules. The result is high-quality synthesis for specific applications with one voice and one style only.

13.7.2 Global Overall Performance

- In this global category fall the overall quality judgments, such as the mean opinion score (MOS), as commonly used in telecommunication applications. Such tests have little diagnostic value, but can clearly indicate whether the speech quality is acceptable for a specific application by the general public. One can think of telecommunication applications such as a spoken weather forecast, or access to e-mail via a spoken output. Also prototypes of reading machines for the visually-impaired, allowing them to listen to a spoken newspaper, are evaluated this way. In field tests not just the speech quality, but also the functionality of the application should be evaluated.

(from http://cslu.cse.ogi.edu/HLTsurvey/ch13node9.html)

"Talking Machines" Human Language Technologies (HLT) Evaluation Workshop, Malta

13.7 Speech Synthesis Evaluation

13.7.3 Towards International Standards

- Although presently there is little standardization and proper multilingual benchmarks for speech synthesis are lacking, various organizations are working on it. Via the Spoken Language Working Group in Eagles, a state-of-the-art report with recommendations on the assessment of speech output systems has been compiled [Eag95], largely based on earlier work within the Esprit-SAM project [PSp92]. The Speech Output Group within the world-wide organization COCOSDA has taken various initiatives with respect to synthesis assessment and the use of databases [PJ94]. One recent intriguing proposal is to arrange real-time access to any operational text-to-speech system via World Wide Web. The ITU-TS recently produced a recommendation about the subjective performance assessment of synthetic speech over the telephone [ITU93,KKSF93].

13.7.4 Future Directions

- In the future, we will probably see more and more integrated text and speech technology in an interactive dialogue system where text-to-speech output is just one of several output options [Pol94a]. The inherent quality of the speech synthesizer should then also be compared against other output devices such as canned natural (manipulated) speech, coded speech, and visual and tactile displays. Also the integration of these various elements then becomes more important, and their performance should be evaluated accordingly.

(from http://cslu.cse.ogi.edu/HLTsurvey/ch13node9.html)

EAGLES

Expert Advisory Groups for
Language Engineering Systems
Spoken Language Working Group

Handbook of Multimodal and Spoken Dialogue Systems

Resources, Terminology and Product Evaluation

Dafydd Gibbon, Inge Mertins, Roger Moore (eds.)

(1997, 2000)

Dedicated to the memory of our colleague, co-author and friend
Christian Benoît

"What is it that is evaluated?".

- **Adequacy evaluations**
  - determine the fitness of a system for a purpose: does it meet the requirements, and if so how well, and at what cost? The requirements are mainly determined by user needs. Therefore user needs have to be identified, which may require considerable effort in itself. Consumer reports are a typical example of adequacy evaluation.
- **Diagnostic evaluations**
  - obtain a profile of system performance with respect to some taxonomy of possible uses of a system. It requires the specification of an appropriate test suite. It is typically used by system developers.
- **Performance evaluations**
  - measure system performance in specific areas. Performance evaluation is only meaningful if a well-defined baseline performance exists, typically a previous version of the system, or a different technology that supports the same functionality. Performance evaluation is typically used by system developers and program managers.

- Three basic components of a performance evaluation have to be defined prior to evaluating a system:
  - **Criterion:** what characteristic or quality are we interested in evaluating (e.g. speed, error rate, accuracy, learning)?
  - **Measure:** by which specific system property do we report system performance for the chosen criterion?
  - **Method:** how do we determine the appropriate value for a given measure and a given system?

(from the Handbook of Multimodal and Spoken Dialogue Systems)

Speech output in multimedia systems

- Taxonomy of output modalities
- Output devices
- Theoretical issues
  - Introduction to multimedia systems
  - Recommendations for the use of speech output in multimedia systems
- Summary of recommendations
  - Recommendations regarding applications
  - Intrinsic properties of speech output
  - Recommendations regarding the environment
  - Recommendations regarding the user
  - Recommendations regarding content
  - Recommendations regarding communicative goals
  - Recommendations regarding interaction
  - Recommendations regarding the combination of speech output and other media

(from the Handbook of Multimodal and Spoken Dialogue Systems, 2000)

Faces and Voices in Action

- speech synthesis represents one modality
  - but increasingly, faces are being included
- there is a need to evaluate not just the speech, but also the contribution of the speech component in an integrated multimodal system for overall communication
- this aspect of evaluation will need action
  - i.e., speech in discourse, non-verbal elements

The Blizzard Challenge 2005

- **In order to better understand different speech synthesis techniques on the same data, we have devised a challenge that will help us better compare research techniques in building corpus-based speech synthesizers.**
- ***The basic challenge is to take the publicly available CMU ARCTIC speech databases and build a synthetic voice. Unknown sentences from an independent source will be generated and each participant will synthesize them with their system. The speech will then be put on the web for evaluation. The results were presented at a special session at Interspeech 2005 -- Eurospeech in Lisboa.***
- The Blizzard Challenge -- 2005: Evaluating Corpus-Based Speech Synthesis on Common Datasets (Alan W. Black, Keiichi Tokuda) PDF
- A Probabilistic Approach to Unit Selection for Corpus-Based Speech Synthesis Shinsuke Sakai, Han Shu PDF
- The Blizzard Challenge 2005 CMU Entry -- A Method for Improving Speech Synthesis Systems (John Kominek, Christina L. Bennett, Brian Langner, Arthur R. Toth) PDF
- Automatic Personal Synthetic Voice Construction (H. Timothy Bunnell, Chris Pennington, Debra Yarrington, John Gray) PDF
- An Overview of Nitech HMM-Based Speech Synthesis System for Blizzard Challenge 2005 (Heiga Zen, Tomoki Toda) PDF
- On Building a Concatenative Speech Synthesis System from the Blizzard Challenge Speech Databases (Wael Hamza, Raimo Bakis, Zhi Wei Shuang, Heiga Zen) PDF
- Multisyn Voices from ARCTIC Data for the Blizzard Challenge (Robert A.J. Clark, Korin Richmond, Simon King) PDF
- Large Scale Evaluation of Corpus-Based Synthesizers: Results and Lessons from the Blizzard Challenge 2005 (Christina L. Bennett) PDF

(from http://festvox.org/blizzard/blizzard2005.html)

CMU_ARCTIC speech synthesis databases

- The CMU_ARCTIC databases were constructed at the Language Technologies Institute at Carnegie Mellon University as phonetically balanced, US English single speaker databases designed for unit selection speech synthesis research.
- The databases consist of around 1150 utterances carefully selected from out-of-copyright texts from Project Gutenberg. The databses include US English male (bdl) and female (slt) speakers (both experienced voice talent) as well as other accented speakers.
- The distributions include 16KHz waveform and simultaneous EGG signals. Full phonetic labelling was performed by the CMU Sphinx using the FestVox based labelling scripts. Complete runnable Festival Voices are included with the database distributions, as examples though better voices can be made by improving labelling etc.

(from http://festvox.org/cmu_arctic/)

towards "better voices"

- is it really just improved labelling?
- what about:
  - prosodic balance?
  - affect & emotional variation?
  - differences in expressiveness?
  - sex, age, and personality differences?
  - conversational speech mannerisms?
  - laughter, grunts, and non-speech noises?
  - mimicry, quoting, acting, whispering, etc., etc.,

towards better speech databases

- not just studio-based 'read-speech'
  but also field-collected 'real-speech'
  - not just from 'experienced voice-talent'
    but also from 'the kid next door'
  - not just 'announcement-style speech'
    but also 'talking for the fun of it'
  - not just 'a few thousand sentences'
    but also 'several year's of talk'

(see for example the ATR-JST/CREST 'Expressive Speech' Corpus)
(http://feast.atr.jp)

synthesis databases

- most current speech databases have been collected for speech recognition research
- synthesis requires:
  - tens or hundreds of hours per speaker
  - variety in speaking styles/expressiveness
  - different interlocutor/speaker-state conditions
  - high-quality recordings (outside the studio)

ECESS

European Center of Excellence on Speech Synthesis

- To achieve the goals of ECESS to push the TTS technology and to speed up the process from basic research to product, clearly defined procedures for evaluation of developed TTS components and assessment of the TTS systems have to be deployed. The main goal in the field of system evaluation will therefore be to establish a common test-bed for evaluation of modules, reference TTS system, and TTS systems developed by the partners. A benchmark test environment that will be developed by the partners, will further contribute to the development of the technology.

- Naturalness, intelligibility, and accuracy are usually evaluated using the subjective measures (based on listening tests). Many elaborated standards and recommendations within different standardisation bodies and groups were set up for defining a framework for subjective evaluation tests and measures, with most widely used MOS listening test, which is most frequently used for evaluation of coded speech. Although human listeners are the ultimate reference also for evaluation of synthesised speech, implementation of subjective tests is usually time-consuming and expensive. Further problem in implementation of listening tests is evaluation of multilingual TTS systems, as for each language the native speakers should be deployed. Different objective measures were developed in the past to compensate in some extend the problems of implementing the subjective tests. Although they can not replace yet the subjective measures in the assessment tests, they are helpful in evaluation of particular component or processing stage in a TTS system.

- As one of the goals of the evaluation is also to build up a multilingual framework for development of TTS systems, the selection of objective evaluation measures that will enable comparability across languages will be an important objectives.

(from http://www.ecess.org/)

ECESS

European Center of Excellence on Speech Synthesis

- On the pragmatic level, the first question is how to determine the proper speaking style in a given dialog situation. So, for instance, when the driver of a car in heavy traffic is listening to the voice of the automatic speaking system it has to sound either as giving a warning, or at least as attracting attention, or also only just as giving some information about the available choices in a given driving situation. The pragmatically adequate speaking style could even have to change within a single utterance. Still on the pragmatic level and depending on the nature of an application even the type of a speaker itself may play a crucial role. One example is that the characteristic voice properties of a synthetic speaker may even be chosen to define the so-called speech logo for corporate identities. On the other hand, the choice of the individual voice quality of a synthetic speaker will mainly have to depend on the nature of the given application. Even the gender and the age of the speaker may have to be taken into account. And it is absolutely true that one cannot use always one and the same speaker for the output of all different man-machine systems. A technical information system such as a telephone directory requires another speaker type than a toy for children, or the agent in a system for adults playing games, or a teaching system for the acquisition of a new second language. If more than one speaker is needed in an automatic dialog system the distinguishability of their voices and their bindings to certain functionalities becomes also a relevant question.

(fromhttp://www.ecess.org/)

communicative goals and interaction

Two channels have been distinguished in human interaction. One conveys messages with a specific semantic content (verbal channel); the other (the non-verbal channel) conveys information related to both the image content of a message and to the general feeling of the speaker. Enormous efforts have been undertaken in the past to understand the verbal channel, whereas the role of non-verbal channel is less well understood.

[...]

To understand non-verbal information, advanced signal processing and analysis techniques have to be applied, and psychological and linguistic analyses must be performed. Moreover, understanding the relationship between the verbal and non-verbal communication modes, and progress towards their modelling, is crucial for implementing a friendly human computer interaction (HCI) that exploits the generation of synthetic agents and sophisticated human-like interfaces.

(from a Preliminary Proposal for a new COST Action)

voice quality control

- from whisper to shout
- from sexy to authoritative
- from gentle to abrupt
- from intimate to formal
- etc., etc.,

- not just the speech, but the voice as an independent variable for evaluation

emotional speech synthesis

- is it really 'emotion'?
  - interest, boredom
  - hesitation, politeness
- anger / sadness / fear / joy . . .
  - fear for the world in today's neocon struggles
  - sadness for the situation in the middle east
  - anger at what is happening
- how do we express these feelings?
- what is 'neutral' ???

human speech communication

- we don't just talk to make announcements
- we also talk to communicate
  - to communicate social relationships
  - to communicate feelings and affect
  - to pass the time enjoyably
  - to establish bonds
  - to joke, etc.,
- and multimodal systems will need equivalent abilities – both in sensing and in synthesis

my recommendations

- I believe we should shift the design of speech synthesis evaluations away from looking at what we CAN do, towards looking at what we CAN'T yet do . . .
- In this way, perhaps we will encourage the designers and producers of these systems to face the challenges of reproducing interactive human speech communication in all its true range of complexities

thank you