Report on the Third ESCA TTS Workshop Evaluation Procedure

Jan P. H. van Santen[1], *Louis C. W. Pols*[2], *Masanobu Abe*[3], *Dan Kahn*[4], *Eric Keller*[5], *and Julie Vonwiller*[6]

[1]Bell Labs – Lucent Technologies, Murray Hill, NJ, USA. [2]University of Amsterdam, The Netherlands;
[3]NTT, Kanagawa, Japan; [4]E-Speech, Gillette, NJ, USA; [5]University of Lausanne, Switzerland;
[6]APPEN Speech Technology, Gordon, NSW, Australia.

ABSTRACT

This paper provides a description and rationale for the Evaluation Procedure taking place at the Workshop. The procedure has three goals. First, setting a precedent of providing conference participants with a more candid and thorough picture of the quality of current TTS systems than is usually available in the form of prepared conference demonstrations. Second, providing results that will be informative for TTS systems developers. Third, stimulating a discussion and contributing to a consensus building process on text-to-speech synthesis evaluation.

1. INTRODUCTION

During previous text-to-speech workshops and speech conferences, the only exposure participants were given to TTS systems was in the form of prepared demonstrations, or brief sessions where a participant could try out some text on a locally installed TTS system. This inevitably led to great uncertainty concerning the true quality of these systems. The current workshop improves significantly on this state of affairs by providing a format whereby text-to-speech systems are presented with the same unknown text materials, covering a variety of text types. The procedure, however, should not be considered as a formal evaluation, because the listeners are the participants themselves and the experimental procedure is necessarily short. To reflect this informality, system-specific results will not be published. We nevertheless believe that this process serves at least three important goals:

1. To give workshop participants a thorough and honest impression of current systems.
2. By exposing participants to a formal listening procedure, to stimulate discussion of evaluation procedures in general.
3. To provide valuable feedback to systems developers and researchers about strong and weak points of their systems.

An attempt will be made to cover as many languages as possible. The limiting factors are: TTS system participation (at least 3 systems are needed to make comparison meaningful for a given language), availability of appropriate textual materials, and availability of enough appropriate listeners. This paper describes the evaluation procedure. For a recent overview of TTS evaluation procedures, we recommend the chapter by van Bezooijen and van Heuven in the EAGLES Handbook [4].

2. ENSURING FAIRNESS

A key concern faced by the organizers was to avoid non-participation due to concerns about fairness and negative publicity. We believe that we have taken significant steps to alleviate these concerns:

- Text materials will mainly be created by standard automated methods, based on text corpora owned by an organization [e.g., Linguistic Data Consortium] that has no formal or informal ties to any TTS system. See [3].
- Selected text materials will be unknown to participating systems.
- System-specific results will not be published, based on the consideration that the subject population – the workshop participants – is not representative of the population at large. Results will only be reported in oral form at the workshop itself.
- The experimental design underlying the listening procedure guarantees that (1) each TTS system will be listened to with the same input text, and (2) a given listener hears different text items on each presentation, thereby avoiding potentially biasing text repetition effects.
- Text materials will cover a range of challenges. Three text types will be used: meaningful text, unpredictable sentences, and telephone directory listings (consisting of names of persons, addresses, and telephone numbers).
- Partial participation is accepted, i.e. participation in only one or two of the three text types.

3. PROCEDURAL DETAILS

3.1. Text Types

Four types of text will be used: Newspaper text (easy vs. difficult), Semantically Unpredictable Sentences, and telephone directory listings. Not all of these types will be available for all languages, and for some languages no text might be available at all.

Newspaper text Two selection methods are used: minimum word frequency based selection, and overall trigram frequency based selection. We explain these methods now.

Easy: Minimum word frequency based selection. This method guarantees that all words in a selected sentence have a frequency of occurrence above some pre-specified threshold. As a result, the sentences will be relatively easy for the grapheme-to-phoneme components of systems. However, in the absence of further constraints, the sentences can have complicated structures (making phrasing and prominence determination difficult, and thereby taxing the prosodic components). For the same reason, they may also contain unusual phoneme sequences at word junctures. This method involves the following steps:

1. For each word in the corpus, determine its frequency (number of occurrences) in that corpus.

2. For each sentence, determine the frequency of the least frequent word.
3. Sort sentences in descending order by least frequent word frequency,
4. Randomly select from the top 1 percent of this sorted list

Difficult: Overall trigram frequency based selection. This method uses successive letter triples as the basic unit. It maximizes the [frequency-weighted] diversity of triphones in the selected text, without concern for word frequency. The selected sentences exercise many components of a TTS system: grapheme to phoneme conversion, acoustic units (in concatenative systems), and prosodic components. It involves the following steps:

1. Determine number of occurrences (frequency) of each trigram in the corpus.
2. For each sentence, add the log frequencies of all its trigrams.
3. Sort sentences in descending order by log frequency sum.
4. Randomly select from the top 1 percent of this sorted list.

Semantically Unpredictable Sentences This procedure has been documented extensively [1, 2] . It involves short, semantically unpredictable sentences of several different, common syntactic structures with words randomly selected from lexicons with frequent, "mini-syllabic" words (smallest words available in a given category. Examples (for English; for other languages appropriate structures are used):

- Subject - Verb - Adverbial
- Subject - Verb - Direct object
- Adverbial - Transitive verb - Direct object (imperative)
- Q-word - Transitive verb - Subject - Direct object
- Subject - Verb - Complex direct object

This test primarily taxes the segmental intelligibility of systems.

Telephone directory listings Pre-formatted telephone directory entries, with the following exact format.

<FirstName> <LastName>, <StreetAddress>, <City>,
<County, Province or State>, <Telno>

Here:

- <FirstName> and <Lastname>: randomly drawn from large lists (with appropriate constraints on frequency)
- <StreetAddress>: fixed format for a given language:

  English: <Number> <Name> Street

  Dutch: <Name>straat <Number>

  French: <Number> rue <Name>

  German: <Name>Strasse <Number>
- <Telno>: provided with appropriate spacing [e.g., US English: aaa-bbb-cccc; Dutch: 0aaa-bb cc dd, or 0aa-bbb cc dd]

Participants were encouraged to produce pre-processing filters for this specific format.

3.2. Listeners

As implied by the above, the listener population consists of participants to the workshop. A key condition for participation in evaluating systems for a given language is that one is sufficiently fluent in that language. Two categories will be allowed to participate: native vs. fluent. Listeners will be identified as such, to allow comparison of results as a function of fluency level. Only if there is time will other listeners be allowed to participate.

3.3. Listening process

Presentation The presentation of an item (a trial) consists of one or more of the following steps:

1. Utterance presented via headphones
2. Rating scales and/or problem areas.
3. Transcription
4. Text presented on the screen
5. Rating scales for correspondence between speech and text

The "Rating scales and/or problem areas" item refers to asking two types of questions. First, ratings (from poor to excellent, on continuous scales) on such dimensions as "overall voice quality" or "naturalness". Second, we identified some key problems, such as "wrong syllable stressed", "bad durations", and "outright mispronunciation". It has been shown elsewhere that non-experts can give consistent answers to questions of this type [5]. Since the current audience consists of TTS experts, we felt all the more confident that these questions were useful. We also include on-screen help describing the problem areas. For example:

> *Wrong Syllable Stressed*: For example, "began" with stress on the first syllable. This also includes the wrong accent (presence/absence, and mora location) in Japanese.

or:

> *Phrase Boundaries*: E.g., a pause after the word "and" in "The blue house and the red house".

At the time this paper was written, certain details have not been finalized yet. These details appear on http://www.itl.atr.co.jp/cocosda/synthesis/evaltext.html

Acoustics Speech files will be played at 11.025 kHz over headphones. Only 11.025 kHz files, 16 bit linear, mono, either .wav or .aiff format are used. The key reason for this restriction is that this is the default setup on most current personal computers.

3.4. Experimental Design

As explained above, the listening procedure consists of a series of trials, where each trial consists of presentation of a synthesized utterance over headphones, and prompts for various responses from the listener. The experimental design has the following properties:

1. Blocked by Text Type within listeners.

2. Same text items for each listener for a given language.
3. Each listener listens to each TTS system equally often
4. Across listeners, each TTS system is presented exactly once with each text item

Standard experimental designs exist that have these properties [5]. The following table explains the design in detail. Systems are denoted S_1, S_2, S_3, ..., S_N; text types T_1, T_2, T_3, ..., and items for text type T_j: $T_j(1)$, $T_j(2)$, ... , $T_j(M_j)$. Thus, the n-th trial for listener L can be denoted $< S_i, T_j(k) >$, meaning that TTS system S_i will be presented with text item k from text type T_j as input. Listeners are grouped in groups of size N, and M_j is a multiple of N. For a given group of N listeners, and text type T_j, the design is constructed by combining several "blocks" having the following structure (rows: trials; columns: listeners):

Once a block is constructed, the trials in each column can be independently randomly reordered. In a typical case, N=7 listeners would be processing N=7 systems, and each would listen to M_1=28 (4 blocks) text items for text type 1, M_2=14 (2 blocks) text items for text type 2, etc., for a total of, say, 84 trials. Of course, for a given set of 7 systems, there can be multiple groups of 7 listeners. The key condition for the above design to be possible is that both the number of listeners and the number of text items are multiples of N=7. Also note that this design allows for the possibility that different groups of listeners for the same set of systems may use different sets of text items.

This design provides unbiased estimates of system performance, but only if we assume that there are no interactions (in the analysis-of-variance sense) between the Listener factor and the System factor. However, this assumption is naive in the light of the fact that systems developers will be among the listeners; even if they attempt to avoid conscious biases in their responses, familiarity with their own system is likely to have a biasing effect.

However, listeners will be asked to reveal their affiliations with systems. This allows us to correct overall scores with two different methods. One is to eliminate from the analysis all responses by a listener on trials involving the system a listener is affiliated with. The other is to assume a simple bias model, where the response of a listener L to system S_i is given by $\alpha_L + x_i$ if listener L is affiliated with system S_i, and by $\beta_L + x_i$ if listener L is not affiliated with system S_i. In either method, it is assumed that listeners affiliated with one system will all exhibit the same bias towards all the other systems. While still naive, this assumption is less likely to be grossly violated than the assumption that there will be no bias in favor of one's own system.

The linear model also allows measuring the bias ($\alpha_L - \beta_L$) for each individual listener. These biases, averaged over the listeners affiliated with a given system, will be made part of the final (oral) report.

3.5. Notification of results

An oral presentation will be given on the last day of the conference on behalf of the 1998 Text-to-Speech Workshop Evaluation Advisory Committee (consisting of the authors of this paper), presenting the results of the evaluation. These results will include the names of the systems. However, we stated as a condition for participating in the workshop the willingness to promise not reveal the identities of participating systems in conjunction with the evaluation results. Thus, one can publish which systems participated, and what ratings the average system obtained, but not which system received which ratings. We decided on these special rules once it became clear that several potential participants were quite concerned about adverse publicity.

4. PARTICIPATING SYSTEMS

Table 2 has a list of the participating systems and the languages or dialects. For many languages, both male and female voices are available. Only for American English and German were the numbers of systems sufficient for both voice genders to allow separate groupings. In all other cases, we requested participants to choose between their male and female voice systems.

5. CONCLUSIONS

As stated in the Introduction, the evaluation process taking place at this conference is not intended as a formal evaluation that would provide conclusive data about the quality of current TTS systems. Instead, our far more modest ambition is to provide the participants with honest demonstrations, where systems are exposed to unknown text of various types, in a setting where listeners can do side-by-side comparisons with other systems. Our hope is that future workshops can build on the lessons learned from the evaluation exercise at the current conference, and that this exercise will significantly contribute to a concensus on TTS evaluation in general.

ACKNOWLEDGMENTS

We thank Rob van Son for the evaluation software, and the Linguistic Data Consortium and NTT for providing text corpora.

6. REFERENCES

1. Benoit, C. An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity. *Speech Communication 9* 1990, 293–304.
2. Benoit, C., Grice, C., and Hazan, V. The SUS test: A method for the asessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication 18* 1996, 381–392.
3. Pols, L., van Santen, J., Abe, M., Kahn, D., and Keller, E. The use of large text corpora for evaluating text-to-speech systems. In *Proceedings First International Conference on Language Resources and Evaluation* (Granada, Spain, May 28–30 1998), vol. 1, pp. 637–640.
4. van Bezooijen, R., and van Heuven, V. Assessment of Synthesis Systems. In *Handbook of standards and resources for spoken language systems*, D. Gibbon, R. Moore, and R. Winsky, Eds. Walter de Gruyter & Co, Berlin, 1998, ch. 12, pp. 481–563.
5. van Santen, J. Perceptual experiments for diagnostic testing of text-to-speech systems. *Computer Speech and Language 7* 1993, 49–100.

Table 1: Building block of the design of the listener experiments. Each cell represents an utterance by system S_i of text item k ($k = 1, \cdots, M_j$) for text type T_j. Rows correspond to trial number ($1, \cdots, N$), columns to listeners. Thus, on Trial 2, Listener 3 hears the second text item of text type T_j, as uttered by System S_4. The total ensemble of presentations to a group of N listeners consists of several of these blocks of N trials each.

Trial Number	Listener 1	Listener 2	Listener 3	... ...	Listener N
1	$S_1T_j(1)$	$S_2T_j(1)$	$S_3T_j(1)$	...	$S_NT_j(1)$
2	$S_2T_j(2)$	$S_3T_j(2)$	$S_4T_j(2)$	...	$S_1T_j(2)$
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
N	$S_NT_j(M_j)$	$S_1T_j(M_j)$	$S_2T_j(M_j)$	...	$S_{N-1}T_j(M_j)$

Table 2: Participating systems

Organization	Country	Language	Organization	Country	Language
ATR-ITL	Japan	US English	TIK/ETH	Switzerland	German
AT&T Research	USA	US English	Tech. U. Dresden	Germany	German
Apple	USA	US English	ETI-Eloquence	USA	Ib. Spanish
BaBel Technologies SA	Belgium	US English	ETSI Telecomunicacion	Spain	Ib. Spanish
CSTR/Festival	UK	US English	Telefonica	Spain	Ib. Spanish
ETI-Eloquence	USA	US English	UPC	Spain	Ib. Spanish
Lucent Technologies Bell Labs	USA	US English	Univ. Politecnica Madrid	Spain	Ib. Spanish
Microsoft	USA	US English	CSELT	Italy	Italian
OGI/Festival	USA	US English	ETI-Eloquence	USA	Italian
Panasonic	USA	US English	Bell Labs	USA	Italian
APPEN Speech Tech.	Australia	Aust. English	ATR-ITL	Japan	Japanese
Telefonica	Spain	Basque	IBM	Japan	Japanese
Telefonica	Spain	Catalan	Bell Labs	USA	Japanese
UPC	Spain	Catalan	Mitsubishi	Japan	Japanese
Fluency Speech Technology	Netherlands	Dutch	NTT	Japan	Japanese
IPO	Netherlands	Dutch	Toshiba	Japan	Japanese
Telia Promotor AB	Sweden	Dutch	University of Tokyo	Japan	Japanese
BaBel Technologies SA	Belgium	French	ATR-ITL	Japan	Korean
British Telecom	UK	French	ETI-Eloquence	USA	Mand. Chinese
ETI-Eloquence	USA	French	Institute of Acoustics/CAS	China	Mand. Chinese
ICP	France	French	Bell Labs	USA	Mand. Chinese
Lucent Technologies Bell Labs	USA	French	Nat. Chiao Tung Univ.	Taiwan	Mand. Chinese
Universite de Provence	France	French	Tsinghua University	China	Mand. Chinese
University of Lausanne	Switzerland	French	ETI-Eloquence	USA	Mex. Spanish
Telefonica	Spain	Galician	Bell Labs	USA	Mex. Spanish
ATR-ITL	Japan	German	OGI/Festival	USA	Mex. Spanish
BaBel Technologies SA	Belgium	German	INESC/CLUL	Portugal	Portuguese
ETI-Eloquence	USA	German	Bell Labs	USA	Romanian
IKP/Bonn	Germany	German	Bell Labs	USA	Russian
INLP/Stuttgart	Germany	German	ATR-ITL	Japan	UK English
Lucent Technologies Bell Labs	USA	German	Aculab PLC	UK	UK English
OGI/Festival	USA	German	British Telecom	UK	UK English
Siemens	Germany	German	CSTR/Festival	UK	UK English
TIK/ETH	Switzerland	German	ETI-Eloquence	USA	UK English