The COCOSDA/LDC Speech Synthesis Evaluation Facilities

Nick Campbell
ATR Interpreting Telecommunications Research Laboratories
Hikari-dai 2-2, Kyoto 619-02, Japan.
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

Abstract

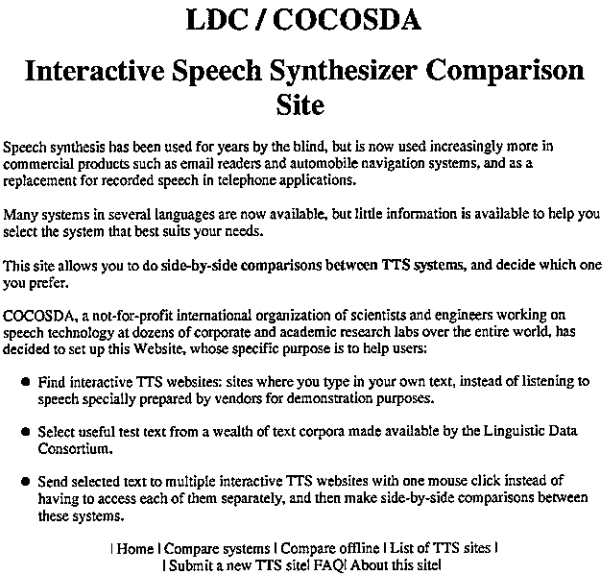
This paper describes the new COCOSDA/LDC Speech Synthesis Evaluation Facilities. The idea for an interactive web site was proposed at the 1997 COCOSDA meeting in Rhodos, Greece, and has now been implemented, ready for the forthcoming Third International Speech Synthesis Workshop which will be held in conjunction with ICSLP-98. The web pages are still in prototype, but have been tested successfully and allow informative comparisons between different speech synthesisers using the same randomly assigned unseen texts.

1 Introduction

One of the three Cocosda Working Groups is concerned with Speech Synthesis Assessment. It was initially coordinated by Louis Pols, and since 1997 by the author. Preliminary discussions for a synthesiser assessment facility were already being held in 1994 within this working group, at the second Synthesis workshop in New Paltz NY, and at the subsequent Cocosda meeting in Yokohama, Japan, where they raised awareness for direct access to TTS websites. Three years later at the Cocosda meeting in Rhodes, Greece, following Eurospeech'97, this possibility was again put forward in a proposal by Jan van Santen and resulted in the formation of the present Cocosda TTS server website committee (see the Acknowledgements section for details).

In its present form the TTS website is not completely finished, but it has shown that by cooperative effort and in a very limited time, a prototype version can be created that allows realistic comparisons of different synthesis methods to be performed. We believe that the potential of this approach has been confirmed, and that it now needs wider use, more contributing sites, and a detailed evaluation from feedback by system designers, providers, and users.

It still has limitations, one of them being that it only allows testing of full throughput rather than component evaluation of individual modules, Another is that there is currently a shortage of non-English text material, but these will be remedied with time. A full test of the facility is planned for the forthcoming Third International Speech Synthesis Workshop [1] which will be held in conjunction with ICSLP-98 at the Jenolan Caves in Australia, where one of the key themes of the workshop will be comparative evaluation of speech synthesis results and methods.



LDC / COCOSDA

Interactive Speech Synthesizer Comparison Site

Speech synthesis has been used for years by the blind, but is now used increasingly more in commercial products such as email readers and automobile navigation systems, and as a replacement for recorded speech in telephone applications.

Many systems in several languages are now available, but little information is available to help you select the system that best suits your needs.

This site allows you to do side-by-side comparisons between TTS systems, and decide which one you prefer.

COCOSDA, a not-for-profit international organization of scientists and engineers working on speech technology at dozens of corporate and academic research labs over the entire world, has decided to set up this Website, whose specific purpose is to help users:

- Find interactive TTS websites: sites where you type in your own text, instead of listening to speech specially prepared by vendors for demonstration purposes.
- Select useful test text from a wealth of text corpora made available by the Linguistic Data Consortium.
- Send selected text to multiple interactive TTS websites with one mouse click instead of having to access each of them separately, and then make side-by-side comparisons between these systems.

| Home | Compare systems | Compare offline | List of TTS sites |
| Submit a new TTS site| FAQ| About this site|

Figure 1: The top page from the LDC/COCOSDA TTS Evaluation web site

2 Details of the site

The Interactive Speech Synthesizer Comparison Site [2] was implemented for initial testing in January this year, and is now stable. At the time of writing (March '98), nine sites had contributed their systems for testing, and five languages (English, German, Japanese, Mandarin, and Spanish) were represented[1]. Contributing sites are listed both by

[1] Ironically, the only Japanese synthesiser yet submitted is from a German University (the University of Duisburg)

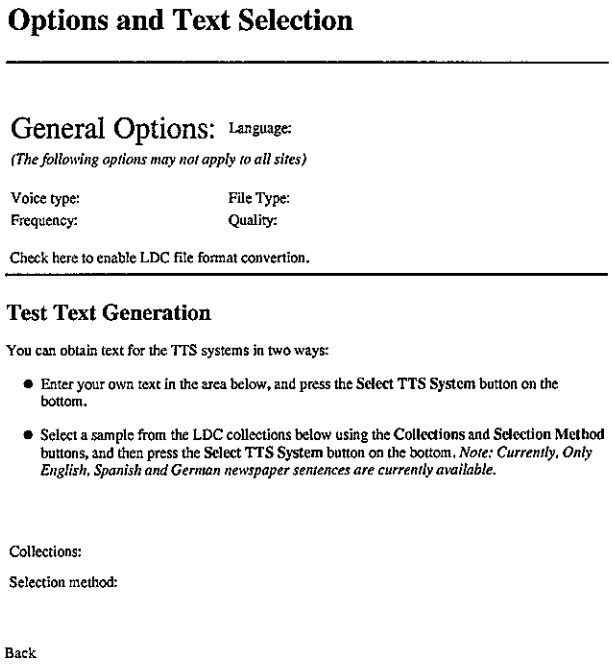


Figure 2: Interactive synthesiser evaluation options

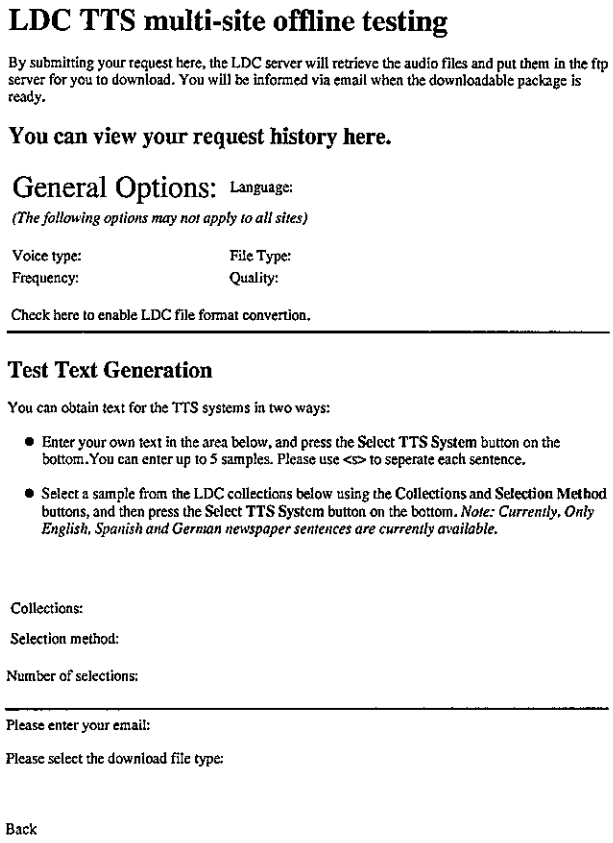


Figure 3: Options for batch-mode evaluation

name and by language for ease of comparison and access.

The LDC/COCOSDA evaluation site offers side-by-side comparisons between TTS systems in order to help potential users decide which voice and synthesis method they might prefer. Previous comparisons have only been available from individual developer sites, many of which only offer pre-stored versions of demonstration speech wave files, which may be difficult to compare, and not necessarily representative of true system performance. By synthesising the same randomly-chosen text through each synthesiser at the site, the listener is offered better opportunities for a fair comparison of the output of each.

2.1 The top page

The top-page menu (figure 1) currently has 7 entries:
- Home
- Compare systems
- Compare offline
- List of TTS sites
- Submit a new TTS site
- FAQ
- About this site

The Home entry provides further background information about the LDC/Cocosda Interactive Speech Synthesizer Comparison Site. Taking the rest in order, "Compare Systems" (figure 2), offers a visitor to the site a selection of voice and text type for utterances to be generated by the various TTS systems. "Compare offline" (figure 3) offers the same facilities, although in this case the audio files are stored for later reference. After the audio samples are retrieved from the individual synthesiser sites, they are compressed and packaged into one big file for ftp. The site currently supports .tar, .gz, .tar, .Z, and .zip formats, offering web-based guidance on each format.

3 The text server

The LDC "text server", generates the novel text for synthesis comparisons, either by rule, or by selection from text corpora, to be sent by the local browser to the CGI server at the interactive web sites and used for synthesis.

3.1 The text selection process

Text for synthesis can be selected by various methods, depending on the needs of the visitor.

Currently the following types are offered: Newspaper sentences, First name/last name combinations, Addresses, Monetary quantities, Dates, Single words, and lists of items of specific tests.

For each text type, the following text selection methods are offered: Random, Minimum word frequency Top 1%, Minimum word frequency Top 5%, Minimum word frequency Top 10%, Maximal Word Coverage Top 1Maximal Word Coverage Top 10%, Overall trigram frequency Top 1%, Overall trigram frequency Top 5%, Overall trigram frequency Top 10%.

The FAQ describes each selection method as follows:

- Random selection.

  This method simply randomly selects a sentence from the entire text corpus.

- Minimum word frequency based selection.

  This method selects short sentences made up entirely of common words. These sentences should pose no problems for most systems. This method involves the following steps:

  1. Determine number of occurrences (frequency) of each word in the text 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, 5, or 10% of this sorted list.

- Overall word frequency based selection.

  This method selects longer sentences with many high-frequency words, although they may contain some rarer words as well. Selected sentences are more taxing.

  This method involves the following steps:
  1. Determine number of occurrences (frequency) of each word in the corpus.
  2. For each sentence, add the log frequencies of all its words.
  3. Sort sentences in descending order by log frequency sum.
  4. Randomly select from the top 1, 5, or 10% of this sorted list.

- Overall trigram frequency based selection. This method uses successive letter triples as the basic unit, but is otherwise the same as overall word frequency based selection. The sentences tend to be long, and may contain several rare words. However, the phoneme combinations tend to be common. Selected sentences tend to be more taxing, in particular for the dictionary and pronunciation rule components of systems.

  This method 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, 5, or 10% of this sorted list.

Submit a new TTS site

Please enter your email address (yourid@yoursite.whatever):

Please enter your web TTS interface URL(http://yoursite.whatever/yourpage):

Please enter the text encoding your TTS supports. if the language has more than one encoding standard such Mandarin (GB, Big5):

<Submit> <Reset>

Back

Figure 4: Instructions for submitting a site

3.2 Submit a new site

This is perhaps the most important page as far as COCOSDA is concerned. We are keen to encourage as many developers as possible to commit their systems to this evaluation methodology. The FAQ offers clear instructions and examples for the CGI and html interfaces required. In principle, the procedure simply requires information about the site address, the interface, and the language encodings for each synthesis system to be included (See figure 4).

3.3 Frequently Asked Questions

The frequently asked questions (FAQ) page is a recent addition which provides answers to such questions as
- what is the goal of this site
- how to set up the browser to listen to speech
- how to set up a web interactive TTS server
- what is an audio format
- what is text encoding
- how is the text selection done
- what is the download file type

As an example, I reproduce below the current entry for setting up a web interactive TTS server

- Write an HTML page to get the text input to the TTS system.

  The easiest way to write the HTML page is to download the sample ⊙. You also can use web pages of other TTS systems as examples. You can modify the example and and put it on your web server.

- Write a CGI program to call the TTS system and send back the audio file.

- What is CGI?

  CGI is not a language. It's a simple protocol that can be used to communicate between Web forms and your program. A CGI script can be written in any language that can read STDIN, write to STDOUT, and read environment variables, i.e. virtually any programming language, including C, Perl, or even shell scripting.

  Here's the typical sequence of steps for a CGI script: Read the user's form input.
  Do what you want with the data.
  Write the HTML response to STDOUT.

- Reading the User's Form Input

  When the user submits the form, your script receives the form data as a set of name-value pairs. The names are what you defined in the INPUT tags (or SELECT or TEXTAREA tags), and the values are whatever the user typed in or selected. (Users can also submit files with forms, but this primer doesn't cover that.)

  This set of name-value pairs is given to you as one long string, which you need to parse. It's not very complicated, and there are plenty of existing routines to do it for you. Here's one in Perl ⊙, a simpler one in Perl ⊙, or one in C ⊙. The CGI directory at Yahoo includes many CGI routines (and pre-written scrpts), in various languages.

  If that's good enough for you, skip to the next section.

- More details:

  If you'd rather do it yourself, or you're just curious, here's the format of the long string:
  "voic=value1&text=value2&audio=value3".
  So just split on the ampersands and equal signs. Then, do two more things to each name and value:

  Convert all "+" characters to spaces, and Convert all "%xx" sequences to the single character whose ascii value is "xx", in hex. For example, convert "%3d" to "=".

  This is needed because the original long string is ¡b¿URL-encoded¡/b¿, to allow for equal signs, ampersands, and so forth in the user's input. So where do you get the long string? That depends on the HTTP method the form was submitted with. For GET submissions, it's in the environment variable QUERY_STRING. For POST submissions, read it from STDIN. The exact number of bytes to read is in the environment variable CONTENT_LENGTH.

- Call TTS system with the text input

  In the CGI script, you can call your TTS system with the parameters specified by the user and have the TTS system generating a temporal audio file.

- Sending an audio back to the user

  First, write the line *Content-Type: audio/x-wav* plus another blank line, to STDOUT. According to audio type requested by the user, you may substitute the *x-wav* to *basic* or *x-aiff*
  After that, write your audio file to STDOUT, and it will be sent to the user when your script is done. That's it. Good luck.

4 What next?

Whereas the present system offers many fine features, there is still plenty of room for improvement.

4.1 Evaluating component modules

The "JEIDA Guideline for Speech Synthesizer Evaluation" (July '95) from the Speech Input/Output Systems Expert Subcommittee of the Committee on Standardization of Office Automation Equipment (The Japan Electronic Industry Development Association (JEIDA)) includes the following:

> There are two types of speech synthesizers, phoneme-to-speech and text-to- speech. Generally speaking, the latter type is much more useful. When using a text-to-speech synthesis system, however, errors occurring in the text analysis part can make it impossible for a user to understand the meaning of the speech message even though the intelligibility of the phoneme-to-speech conversion part is quite high. Therefore, the evaluation for the text analysis part of text-to-speech synthesis system is as important as [the] intelligibility test.

Whereas some might argue with their definition of 'usefulness', the need for a componential evaluation of text-to-spech system modules is clearly stated.

The current evaluation setup makes the implicit assumption that all synthesisers are text-to-speech synthesisers, and that all speech to be produced by them is of the 'read-speech' variety. There is at the moment no facility for testing interactive speech synthesis such as would be required for interpreted communications, nor is there any way of annotating the input text with more appropriate ways of rendering it, other than those produced by the default text-analysis routines of the individual synthesisers.

4.2 Simple words in single sentences

As Pols et al [3] point out,

> The minimum word frequency based procedure, for instance, selects short sentences made up entirely of common words. First, for each word in the text corpus its frequency of occurrence is determined, then, for each sentence the least frequent word is found. All sentences are then sorted according to this least frequent word frequency. Finally a sentence is randomly selected from the top 1, 5, or 10% (specified at input) of this sorted list. This leads to (top 1%) sentences like: "Officials from the federal agencies would not comment". "A 5 percent tax increase was approved last month". The overall trigram based procedure (letter triples) leads to longer sentences containing several rarer words. An example of such a top 1% sentence is: "Hotels negotiate varying contracts with different communications companies, sometimes using one for local calls and another for long-distance service". With this capability one could for instance have several sentences generated by one system, or one and the same sentence by several systems for comparison.

Single sentences such as these provide a good test of how any given text-to-speech system can handle the text-to-phoneme conversions for various input types, and how they predict phrasing and intonation for read speech, but they do not provide an indication of how a synthesiser will sound under repeated use when more sentences from a continuous text are to be synthesised, for example when reading a story. Such paragraph-level intonation prediction is still difficult for many speech synthesisers, and whereas the present system allows good evaluation of phonemic realisation and local prosodic effects, it offers little for the evaluation of the global prosodic effects that make connected text synthesis not just more intelligible, but also less tiring to listen to.

4.3 Providing feedback

In encouraging other synthesis developers to add their systems to the evaluator, it would be useful to offer subjective feedback from the listeners who make use of the site, perhaps by allowing a visitor to complete and return a form such as the following suggested by JEIDA:

> APPENDIX B [Descriptive words for the speech quality] User responses to and impressions of the synthesized speech are evaluated through the use of feature descriptive words (Semantic differential method). The words must be daily-use, familiar and easily understandable ones. Provided below is a list of examples of feature descriptive words used to evaluate the quality of synthesized speech. However, the significance test of these words should be conducted after evaluation since it is not always certain that these words will be suitable as evaluation terminology. Each descriptive word is paired by its antonym. The bipolar rating scales (semantic scales) are composed by using these paired words.
>
> 1. Descriptive Words for the Intelligibility: easy / hard to understand, easily misread / hardly misread.
> 2. Descriptive Words for the Sound Quality: beautiful / dirty, smooth / rough, glossy / lifeless, sharp / dull, full of life / nasal, articulate / muffled, thick / thin, powerful / weak, rich / poor, grave / light, sweet / metallic sound, soft and full / harsh, bright / somber, soft / hard, clear / turbid.
> 3. Descriptive Words for the Temporal Factors: natural / unnatural rhythm, fast / slow, continuous / choppy.
> 4. Descriptive Words for the Intonation: natural / unnatural intonation, natural / unnatural accent, fluent / halting.
> 5. Descriptive Words for the Overall Goodness: human-like / artificial, preferable / unpreferable, excellent / poor.

6. Descriptive Words for the Suitability: easy / hard to hear, comfortable / frustrating, pleasant / annoying, Japanese / foreign, male / female, high voice / low voice, young / old, suitable / unsuitable for the purpose.

5 Conclusion

A significant step has been taken towards facilitating the evaluation of speech synthesis systems. The interactive web-based text server provided by the LDC in conjunction with COCOSDA will undoubtedly prompt more synthesis developers to link their sites and to offer their systems for comparative analysis in the future. In the case of commercial systems, this will allow the potential customer greater freedom of choice before making a purchase. For developers, it will allow them to compare their progress with that of other similar systems.

The site is billed as 'multi-lingual', but there is a not-unsurprising preponderance of English and European texts. It is hoped that this will become more balanced and international as other nations contribute either text corpora or similar text servers.

The texts offered are currently strongly biased towards newspaper reading, thus reinforcing the view of a speech synthesiser as a 'reading machine', rather than as a 'voice interface' for information access systems. It is hoped that with time we will also see the introduction of a facility to test the synthesisers as 'speaking machines', encouraging the evaluation of expression and liveliness of intonation that cannot yet be predicted from written text alone.

Acknowledgements

The idea for this site was proposed by Jan van Santen at the 1997 COCOSDA meeting in Rhodos, Greece. The site was designed by Masanobu Abe, Alan Black, David House, Louis Pols, Jan van Santen, and Mark Liberman, and was implemented by Zhibiao Wu at LDC. On behalf of the COCOSDA Speech Synthesis Working Group, I would like Mark Liberman for kindly donating LDC facilities to this not-for-profit endeavour, and for making the various LDC texts available.

References

[1] The workshop : http:/www.itl.atr.co.jp/cocosda/synthesis/3rd_ws.html

[2] The website : http:/www.ldc.upenn.edu/ltts

[3] Louis C.W. Pols, Jan P.H. van Santen, Masanobu Abe, Alan Black, and David House, "Easy Access via a TTS Website to Mono- and Multilingual Text-to-Speech Systems", in Proc ELRA in Wonderland (1998).