



Spoken Language Synthesis:

Experiments in synthesis of spontaneous monologues.

Shiva Sundaram, Shrikanth Narayanan.

Department of Electrical Engineering,
And Integrated Media Systems Center

University of Southern California, Los Angeles.

ssundara@usc.edu

shri@sipi.usc.edu



What is TTS synthesis?

- ✿ TTS is a text to utterance mapper: a mechanism to convert a given text, to speech utterances that constitute the language in the text.
- ✿ Intelligibility and naturalness are essential factors.
- ✿ Present technology has good speech quality, and intelligibility.



Limitations of a Conventional TTS synthesizer:

- ✱ Intended design : to mimic “reading out of a book”. for data-retrieval, and interaction that last only for few turns (short duration)
- ✱ Lacks “naturalness” : it does not capture all the dimensions of natural speech for the sake of intelligibility.
- ✱ There is no provision for customizable speaking styles: it is not *adaptable*.
- ✱ Not ideal for interactive technologies.



Research in this paper:

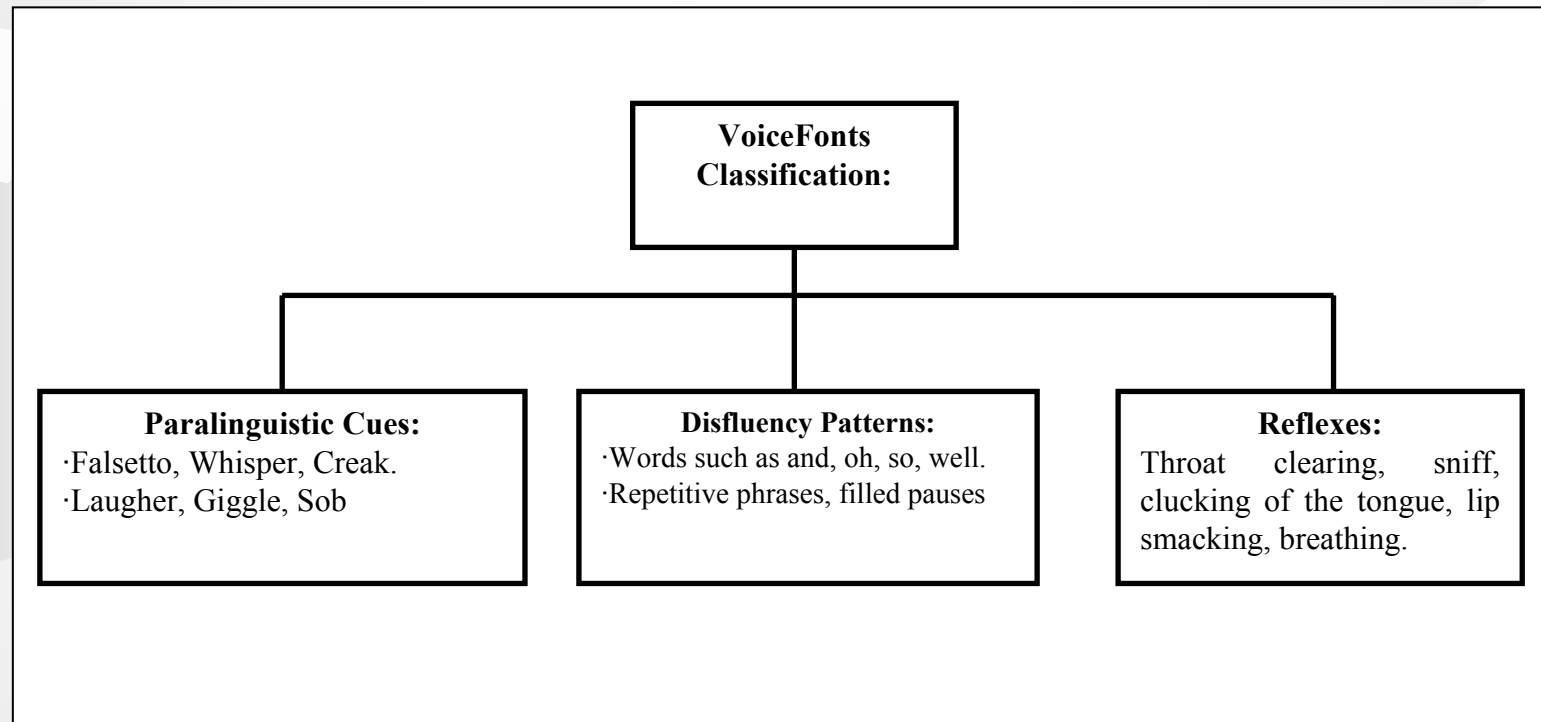
- ✿ Improve naturalness!!
- ✿ Understand and capture an important feature of real speech : **VoiceFonts**.
- ✿ Propose and implement a data driven approach to include these features.
- ✿ Develop techniques to make an adaptable TTS synthesizer. Designed for interactive applications such as:
SUIs, Computer generated Avatars etc.



What are VoiceFonts? 😊

- Usually they are Utterances not part of a language/structure :
 - laughter, tongue clucking, breathing, throat clearing, discourse/disfluency markers.
 - They occur frequently in real spontaneous speech, speaking style of a person depends on their occurrence and usage.

A more formal picture of VoiceFonts:



the features included in this research are:

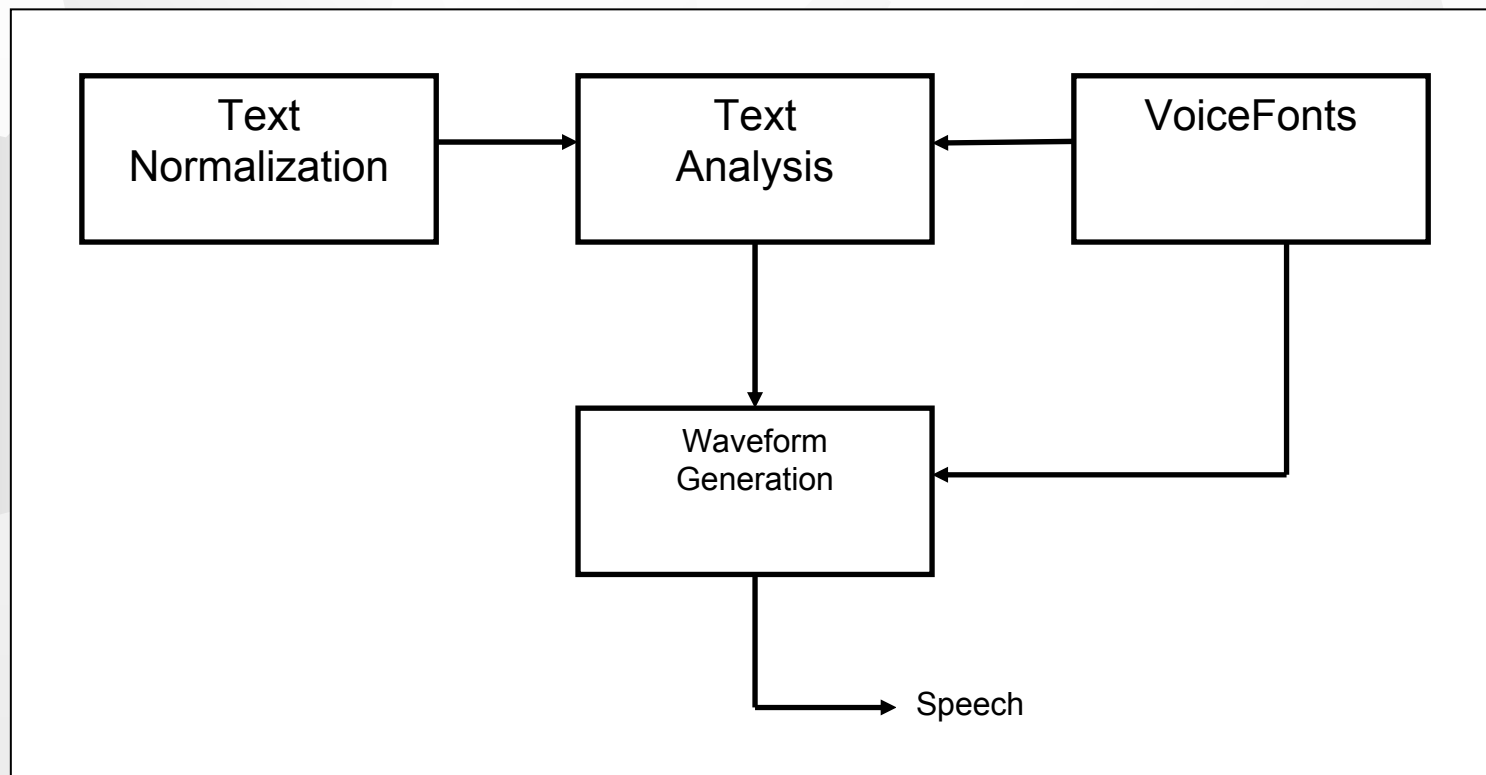
- Laughter and giggle, both as a single category: laughter.
- Breathe in, breathe out, and lip smacking.
- Filled pauses: um and uh, and fillers: and, oh, so, well.



The Spoken Language Synthesizer: technique.

- ★ Study the occurrences of the VoiceFonts and their relative frequencies (bigram probabilities) from the training corpus.
- ★ Generate models for **speaker independent**, **speaker dependent** and **speaker-adaptable synthesis**.
- ★ Include these models in a TTS synthesizer.

The Spoken Language Synthesizer :



Data analysis and VoiceFonts Models:

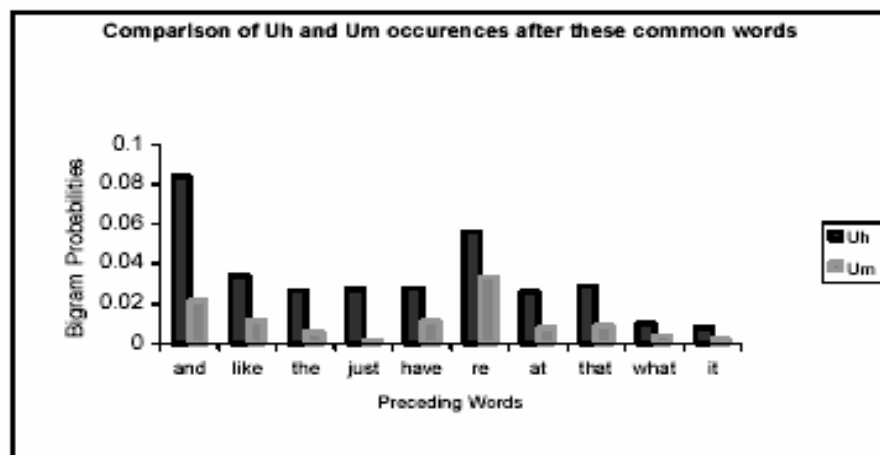


Figure 2: Distribution in SWITCHBOARD corpus

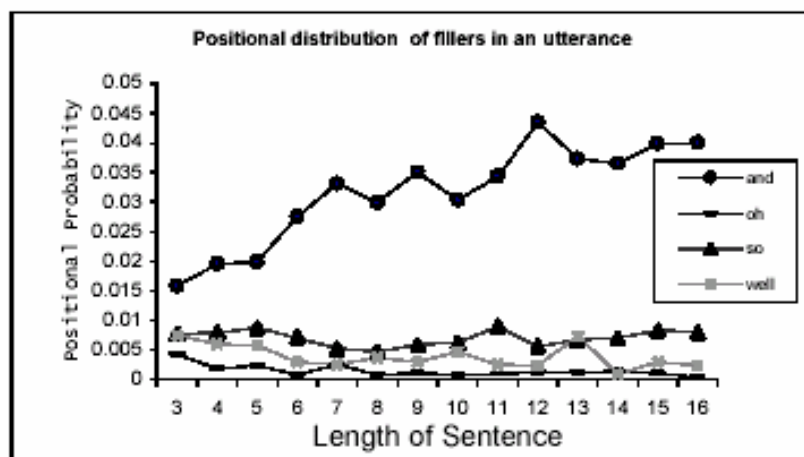


Figure 4: Positional distribution of fillers

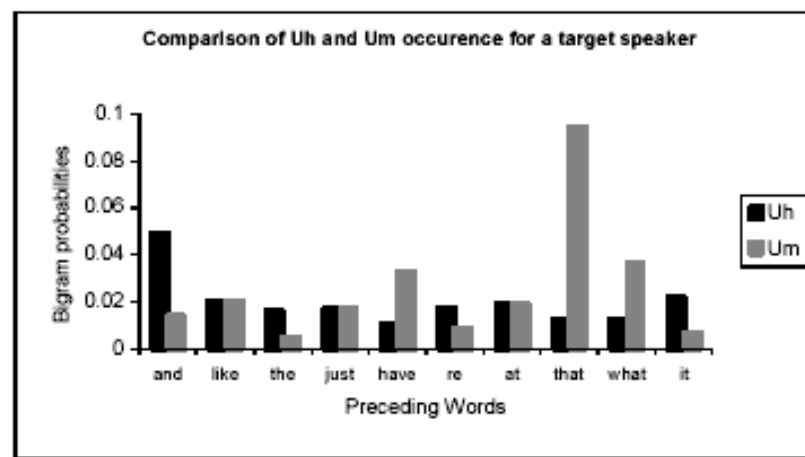


Figure 3: Distribution for a target speaker monologue

Synthesis, LDS :

- ✱ A Limited Domain Synthesizer (LDS) was set up using FESTVOX.
- ✱ For Synthesis, the VoiceFonts were tagged with unique word level symbols.

INPIYT: □Αψεσηα? Σεε τηατ βοοκ ον τηε ταβλε?
Χαν ψου βρινγ ιτ ηερε?□

Τρανσφορμεδ–INPIYT: □Αψεσηα [ΠΑΥΣΕ] σεε
τηατ βοοκ ον τηε [υη] ταβλε [ΣΗΟΡΤ ΠΑΥΣΕ]
[BREATHE IN] χαν ψου βρινγ ιτ ηερε□.

Assuming the units corresponding to the VoiceFonts are present in the synthesis inventory.



Synthesis, LDS : the experiment

- ★ 7 sets of synthesized sentences with 4 types each :
 - ★ Conventional TTS, LDS without VoiceFonts, LDS with VoiceFonts, Original Speech.
- ★ Nineteen volunteers evaluated each clip on a scale of 1 to 5 in terms of 4 subjective qualities: Naturalness, Spontaneity, Fluency and Intelligibility.
- ★ they were also asked to guess the original speech clip at the end of each of the 7 sets, as a 4-way classificatory evaluation.

Results :

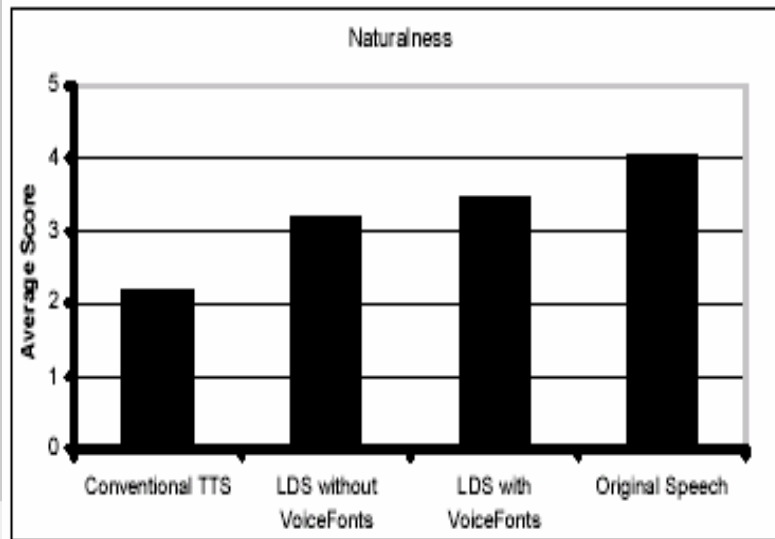


Figure 6: Average Naturalness scores for the 4 categories

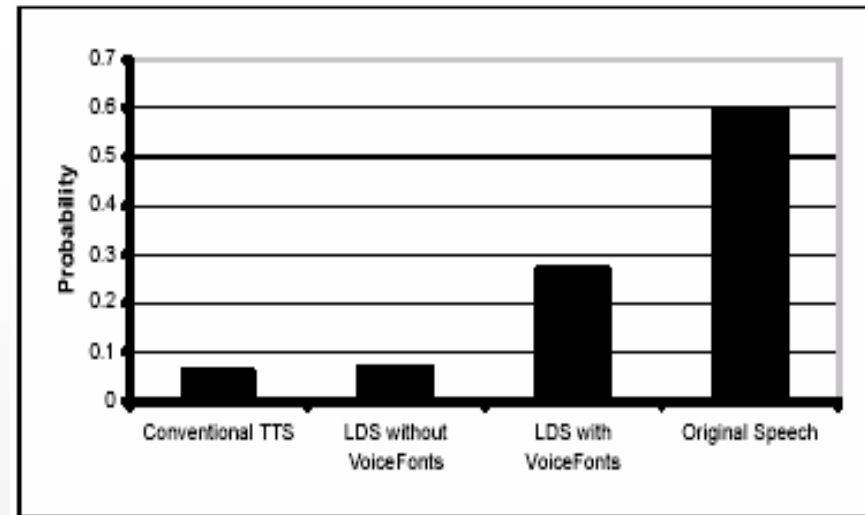


Figure 5: Probability of guessing a clip to be original speech from among 4 choices across 19 subjects.

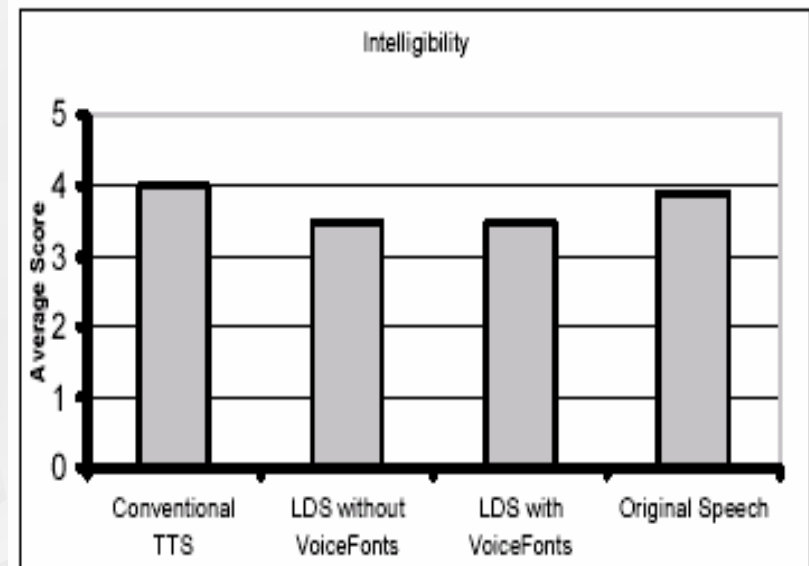


Figure 6: Average Intelligibility scores.




Discussions:

- ✱ Available styles for synthesis would depend on the different training corpus.
- ✱ The inventory of VoiceFonts required would depend on the *style*.
- ✱ The SLS can be used as a better reading machine and for interactive technologies, making it ***adaptable***.



Problems, Limitations and Future Work:

- ✿ In this paper, the language generation problem was not discussed: A direction for future work.
- ✿ The input to the synthesizer was handpicked and not generated automatically.
- ✿ Corpus for different domains of styles were not available.
- ✿ Comparison with conventional TTS is not absolute, since different synthesis techniques were adopted.

- 
- ✿ Evaluation remains difficult, The work presented here deals with synthesis in a different domain than which exists presently.

If the question was to evaluate real speech of people, how would it be done?

Thank You!