SPOKEN LANGUAGE SYNTHESIS: EXPERIMENTS IN SYNTHESIS OF SPONTANEOUS MONOLOGUES

Shiva Sundaram and Shrikanth Narayanan
Department of Electrical Engineering and Integrated Media Systems Center
University of Southern California, Los Angeles, CA 90089
ssundara@usc.edu, shri@sipi.usc.edu

ABSTRACT

While TTS technology has come a long way, there is an ongoing need for bringing improved "naturalness" to synthesized speech. One predominant aspect of natural, spontaneous speech is the variability in it along several dimensions -- in terms of vocabulary, prosodic features, paralinguistic elements and discourse markers. Such variability is typically carefully avoided or minimized in conventional text to speech for the sake of high intelligibility. However, in applications requiring immersive anthropomorphic human-machine interfaces, including those with computer-generated avatars, there is a great desire to mimic human-like synthesized speech output. In this paper we investigate methods and the usefulness of incorporating certain features characterizing fluent natural speech for increasing "naturalness" in synthesized speech. We propose a data driven approach for modeling both speaker-independent and speaker-dependent spontaneous speech features at the lexical and acoustic levels (so-called, VoiceFonts). This method has the potential to create unique, custom *speaking styles* of a target speaker. A simple limited domain synthesizer was built based on this idea using data from a classroom lecture and was used to synthesize 28 target utterances. Results from preliminary listening experiments by 19 volunteers showed that such an approach indeed improves naturalness, without significant loss in intelligibility, beyond the limitations of the underlying waveform synthesis. For example, subjects could correctly identify natural speech with a probability of 0.6 and confused the clips synthesized in this work with natural speech with a probability of 0.27 in a 4-way choice listening test.

1. INTRODUCTION

Data-driven unit selection text-to-speech synthesis (TTS) schemes have contributed toward more natural and intelligible synthesis speech quality. Specifically, approaches where the domain of the text material could be restricted have been shown to produce fairly successful results even with different speaking styles [6]. Although these methods have helped in minimizing artifacts in synthesized speech, even a casual listener could often discern that the "source" is not natural (machine-generated). A significant amount of TTS research focus hence is devoted toward making machine generated speech more natural at various levels -- acoustic, lexical and discourse. "VoiceFonts" [1] in natural speech characterize a speaker and provide a way to discern spontaneous natural speech from prepared or read speech, even those produced by the same speaker. Variability in speech, exemplified by VoiceFonts in this paper (Sec 1.1), occurs at acoustic, lexical and discourse levels, exhibiting both speaker-specific and speaker-independent patterns. Many of these characteristics could be modeled through careful data analysis for later inclusion during synthesis (in this paper, the term synthesis is used broadly to include generation as well). This paper implements methods for including VoiceFont features in synthesizing natural speech with a specific focus on lecture monologues. In a sense, the approach is a reverse of what is adopted in automatic recognition of spontaneous speech where filler acoustic and language models are used to ``filter" some of the VoiceFont effects to improve robustness.

1.1 VoiceFonts

In [1], Campbell introduced the notion of including VoiceFonts in synthesized speech. Some features discussed as VoiceFonts in [1] are: Laughter, tongue clucking, lip smacking, tutting, and inhalation of breath. However the scope of VoiceFonts in spontaneous natural speech could be expanded by the inclusion of additional features that can be categorized into the following:

- Paralinguistic Cues:*falsetto, whisper, creak, laughter, giggle, cry, sob* etc.
- Disfluency patterns: words such as: *and, oh, so well, okay,* repetitions and filled pauses: *uh* and *um.*
- Reflexes*: Throat clearing, sniff/gulp. clucking of the tongue, lip smacking,* and *breathing in.*

Falsetto, whisper, and creakiness are voice quality features and involve prosodic modification of speech. Laughter, giggle, and cry patterns are voice qualifications and their occurrences can be modeled from a statistical point of view for generation. Due to corpus limitations, only laughter and giggle have been included in this study.

Filler words and other disfluency markers do not exactly fall into the category of "VoiceFonts" but are important for studying and synthesizing spontaneous speech. Words such as *and, oh, so* and *well* provide emphasis in the meaning conveyed, and strongly influence the use of filled pauses and the reflexes. Such discourse characteristics in spontaneous speech have been studied before especially from an ASR point of view [2][3].

Reflexes are usually involuntary and are sometimes used by a speaker to "make a point", indicate the beginning of a sentence, or signal the introduction of a new idea. For example tongue clucking usually indicates disagreement. They also affect the fluency of speech. Reflexes and their usage have also been studied in detail [4] and [5].

To limit the scope of this work, the features included here 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.*

Also included are repetitions and occurrences of common phrases such as: *you know, so basically*: which are indicators of disfluencies. To obtain insights into speaker-independent patterns in these VoiceFont features, a transcribed portion of the

SWITCHBOARD corpus was analyzed. The target synthesis domain was a speaker-specific lecture monologue for a native speaker of American English.

1.2. Evaluation and Definitions

A preliminary subjective performance evaluation between the spoken language synthesizer output and original speech was done. The factors included in the evaluation are: Naturalness, Spontaneity, Fluency and Intelligibility. Naturalness of a synthesized speech clip is defined as the closeness between a real speech and the equivalent synthesized output. Spontaneity refers to the *unpreparedness* in context. Fluency and Intelligibility, considered secondary factors for this study, affect the *clarity of speech.* Four different types of outputs of the same target sentences were synthesized and compared:

1. The original sound clips from the training data.
2. Limited-domain Synthesis Output without VoiceFonts.
3. Limited-domain Synthesizer Output with VoiceFonts.
4. Generic (non-spontaneous) TTS: Public-domain AT&T Natural Voices Synthesizer Output [9].

Section 2 describes our method, section 3, the evaluation experiments and results and section 4 provides a discussion.

2. METHOD

The method followed in this paper is simple and the block diagram of the scheme is shown in Figure 1.

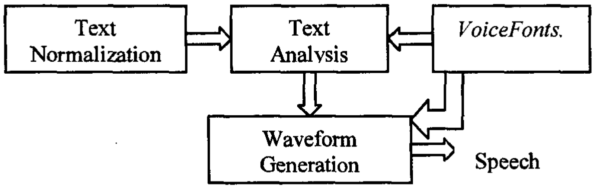

Figure 1: A system for synthesizing spontaneous speech

Text normalization, text analysis and waveform generation are the typical steps involved in a conventional concatenative TTS system. The key block introduced is the Features/models block, which relates to the inclusion of VoiceFonts. Specifically, it includes bigram/unigram probabilities of occurrences of VoiceFonts estimated from the training data (for use in text analysis) and acoustic unit models for voicefonts built from the data (for use in waveform synthesis).

Three spoken language synthesis scenarios are possible:

- Speaker-independent synthesis: the models of VoiceFonts using data statistics obtained from a wide range of speakers are included in the synthesizer.
- Speaker-dependent synthesis: the models of VoiceFonts are specific to a target speaker.
- Speaker-adapted synthesis: the generic voice font models are adapted to speaker-specific data and for specific application domains (such as monologues, dialogues and read speech).

It is important to note that spontaneous speech, unlike written text or prepared speech, is not well structured e.g., in general, it does not have a clear beginning or end to a sentence and often consists of short spans of phrases that make contextual sense but may or may not be grammatically correct. This poses a challenge during transcription and statistical 'language' modeling. If {Vs} is the set of all the words in the speaker's vocabulary and {VF} is the set of all the VoiceFonts used in a particular speaking style, a distribution of {VF} in {Vs} would generate the necessary target speaking style. Unigram probabilities define the set {Vs} and {VF}, and bigram probabilities give an idea of distribution {VF} in {Vs}.

Given a statistical model for VoiceFonts derived from an annotated corpus, the text can be re-annotated for synthesis such as in following example:

INPUT: "Ayesha? See that book on the table? Can you bring it here?"

Transformed-INPUT: "Ayesha [PAUSE] see that book on the [uh] table [SHORT PAUSE] [BREATHE IN] can you bring it here".

Assuming that units corresponding to these voicefonts have been included in the synthesis inventory, the transformed input above could be utilized for waveform synthesis. The transformation of the input primarily deals with language generation. Since the generation is probabilistic, some constrained optimization scheme is required to cull the actual utterance for synthesis automatically from the possible candidates. Such a scheme is a topic for future work. For the experiments in this paper, the target sentences were manually selected from .the list of generated possibilities.

To test the final results, a limited domain synthesizer was set up using the FESTVOX synthesizer [8] (see section 2.2).

2.1. Data Analysis and Modeling

This section describes the statistical analysis of VoiceFonts in transcribed speech. Analysis was done using the CMU-Cambridge Statistical Language modeling Toolkit [8].

For data analysis, two different sets of transcriptions were used: SWITCHBOARD-1 Telephone Speech Corpus (SWB), and 85 minutes of a class lecture from a native speaker of American English recorded at USC.

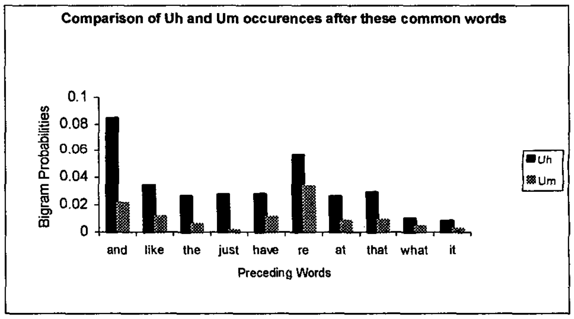

Figure 2: Distribution in SWITCHBOARD corpus

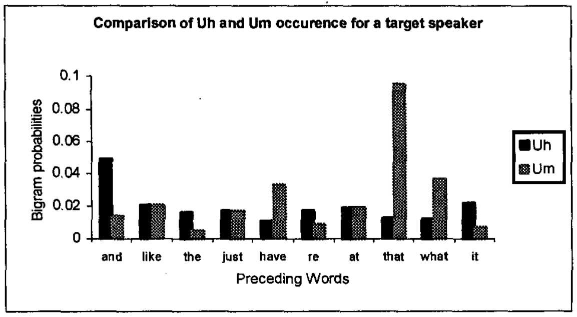

Figure 3: Distribution for a target speaker monologue

The SWITCHBOARD corpus did not include transcriptions of many VoiceFonts discussed previously; nevertheless, it was helpful to understand the usage of certain VoiceFonts over a range of speakers. The lecture speech, transcribed with the extended list of voicefonts, was used to model occurrences of voicefonts for speaker-specific characteristics.

For example, from Figure 2 it can be seen that amongst the filled pauses, the occurrence of uh's was three times the occurrence of um's in the SWITCHBOARD dialog corpus, and were commonly used after the words: *and, the, that, have, like, what, at.* However, the speaker-specific distribution in the monologue was different as shown in Figure 3. Here, the distributions of the two filled pauses are found to be similar. This indicates that usage of VoiceFonts is possibly speaker-dependent (influence of discourse modality is not known here).

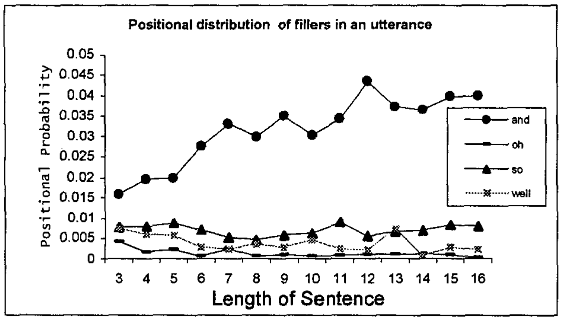

Figure 4: Positional distribution of fillers

W2	w1	And	Oh	So	Well
w2	and	0.0417	0.0080	0.0088	0.0155
	oh	0.0000	0.0134	0.0000	0.0015
	so	0.0279	0.0107	0.0077	0.0046
	well	0.00212	0.0428	0.0055	0.0031

Table 1: Bigram Probability Matrix of SWITCHBOARD discourse Markers.

Figure 4 indicates the positional variation in an utterance of the common fillers in the SWITCHBOARD corpus. The first and second positions are omitted because all the words had similar trends of very high occurrence in the first position and a large drop in the second. An increasing trend in *And* occurrence opposed to a decreasing trend in other markers: *Oh, So* and *Well.* An average length of 16 words was chosen for analysis.

Table 1 indicates the bigram probabilities of the common fillers. For example, the occurrence of "*and oh*" is zero whereas the occurrence of "*oh and*" is 0.0080. Table 1 also provides information about repetitions of these words.

Information such as in Figure 4 and Table 1 are essential during the generation stage to convert a given input text into a more natural spontaneous form, that *would have* been generated by the target speaker, based on the corpus analysis.

2.2. Synthesis

A limited domain synthesizer (LDS) was set up using the FESTVOX synthesizer [8]. Fifty lecture monologue utterances, each about 10-12 words long, were used as the training corpus. The standard procedure for setting up an LDS was followed as described in the FESTVOX manual [8]. For synthesis units, the VoiceFonts were tagged using unique word-level symbols. Once these unique words are inserted as part of the input, the synthesizer would generate the corresponding VoiceFonts during the synthesis.

3. EXPERIMENTS AND RESULTS

Seven sets of utterances were used for the evaluation experiments. Each set had one utterance for each of the 4 categories compared in this paper and presented in a random fashion: a generic unit selection TTS for "non-spontaneous" speech (using the AT&T NaturalVoices Synthesizer), an LDS without VoiceFonts, LDS with VoiceFonts, and the original speech clips from the training data. Nineteen volunteers evaluated each clip on a scale of 1 to 5 (1-Very Poor; 2-Poor; 3-Average; 4-Good; 5-Very Good) in terms of 4 subjective qualities: Naturalness, Spontaneity, Fluency and Intelligibility. The participants were asked to guess the original speech clip at the end of each of the 7 sets, as a 4-way classificatory evaluation.

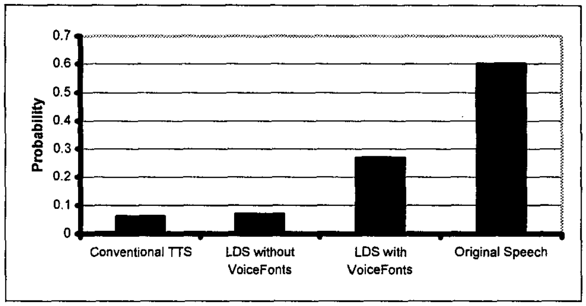

Figure 5: Probability of guessing a clip to be original speech from among 4 choices across 19 subjects.

Figure 5 shows the probability of an utterance in each of the set to be guessed as the original speech clip. The clip directly from the training corpus could be guessed correctly as the original speech with a probability of 0.6. The synthesized utterances with the VoiceFonts were guessed to be the original with a probability of 0.27; the generic synthesizer and the LDS without voice fonts had equal low probabilities of confusion of about 0.08. This indicates that the LDS outputs with VoiceFonts has 3 times the probability to be confused as the original speech clip, than the LDS without VoiceFonts, which can be seen as a direct indication of improvement due to the addition of VoiceFonts. Out of 19 participants only 4 could correctly guess the original speech clip in all of the 7 sets. It should be noted that avoidance of the perceivable artifacts in the waveform generation of our LDS implementation could have contributed to a smaller difference than the 0.6 and 0.27 probability of classification between the original speech and LDS with VoiceFonts.

A single factor ANOVA was performed for each of the 4 evaluation categories. Scores for naturalness and spontaneity were significantly different across the 4 experimental categories with $p \leq 0.001$and for fluency with $p \leq 0.01$. Differences in intelligibility scores were not found to be statistically significant across the four conditions.

Figure 6 indicates the advantage in naturalness of Limited Domain Synthesis over domain-independent concatenative synthesizers, an expected result. Comparing within the LDS outputs, the clips with VoiceFonts had a higher naturalness score of 3.45 as to the average score of 3.17 for the LDS clips without VoiceFonts. Original speech had the highest average score of 4.0. Results were statistically significant ($p \leq 0.001$).

In the evaluation, the subjects were also asked to write down what they considered signifying spontaneity in speech. Almost all the participants mentioned that spontaneity is a measure of unpreparedness in speech. This in turn has implications for the language generation problem. It should be noted that the spontaneity scores here are somewhat limited in their scope since there was no discourse context information for the synthesized speech of our experiments.

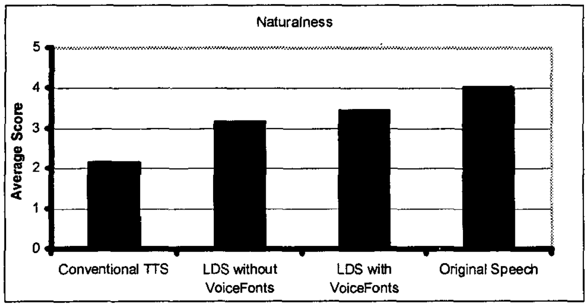

Figure 6: Average Naturalness scores for the 4 categories

However, results indicated that the average spontaneity scores for both types of LDS outputs were similar, and were slightly higher than the generic-domain independent synthesizer. The original speech clip had the highest average score of 3.94. The standard deviation in the scores for LDS was smaller (0.78) than for the original and generic TTS cases (0.83).

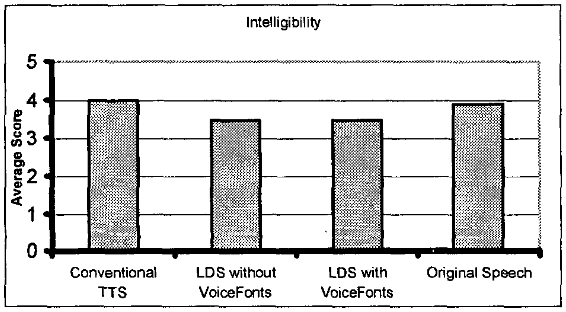

Figure 6: Average Intelligibility scores.

In Figure 7 there is a decrease in intelligibility, albeit not statistically significant, for the LDS schemes as compared to original speech. One reason was due to the presence of artifacts in our synthesized clips; the decrease is also a price paid for increase in naturalness through the use of units derived from spontaneous speech for synthesis. For example, in a real lecture, unless the lecturer is completely prepared (not spontaneous), he would frequently use filled pauses and repeat certain phrases, causing a reduction in intelligibility. Compare this with the results for the generic TTS, which represents a "cleaned-up" version without voicefonts. Fluency can be considered to depend on two factors: synthesis quality and language flow; both are dependent on the inclusion of VoiceFonts. Interestingly, the scores for fluency were similar for the conventional TTS, and LDS with and without voicefonts, but significantly higher for original speech clips.

4. DISCUSSIONS AND CONCLUSIONS

This paper presented an idea for extending the capability of current TTS systems to spontaneous speech for increased naturalness in synthesized speech quality. The synthesizer in this work belongs to a class that can generate spontaneous speech, specifically adapted to a real-speaker. The listening tests indicate that including VoiceFonts in synthesized speech improves its perceived naturalness, moving it closer to real speech. The naturalness is higher than for a generic synthesizer that targets natural, highly intelligible *non-spontaneous* speech without any significant loss in intelligibility. Like any TTS system, the spoken language synthesizer could be used as a better reading machine used in applications such as human-computer interaction, computer generated avatars, and SUIs (Speech User Interface) where natural spontaneous speech and customized styles are desirable.

The work presented here deals with synthesis in a different domain than which exists presently, posing additional challenges for evaluation. If the question were to evaluate real speech of people, how could it be done? Speaking style, discourse context and the application domain strongly affect these evaluation metrics. For example, in public speaking, a good speaker would have minimal of filled and silent pauses, repetitions and discourse markers, whereas in normal conversational speech, filled pauses and discourse markers are prevalent and are characteristic of a natural dialogue interaction between humans. This brings up the question of *suitability* and *adaptability* of a system. State-of-the-art TTS systems perform acceptably well in applications such as data retrieval, which last only for a few turns (short duration) but are highly unsuitable for long-term interactions such as immersive SUIs. For example, in computer-generated avatars, the entity may be required to deliver a monologue or simply converse with the user. The approach used in this paper could be used to customize speaking styles specific to the demands of diverse user environments. The results are promising, indicating that naturalness could be improved by including VoiceFonts and customized language generation schemes. There are however clear limitations of the work including those due to the LDS implementation resulting in possibly confounding synthesis artifacts and the lack of context in presenting stimuli (monolog, dialog) for subjective tests. Further, an optimized VoiceFont generation scheme is lacking. These are some of the directions for future work.

5. REFERENCES

[1] N. Campbell, Where is the information in speech? (And to what extent can it be modeled in synthesis?), Jenolan Caves TTS workshop, 1998.

[2] E. Shriberg, Disfluencies in SWITCHBOARD Speech Technology Research Lab. SRI international.

[3] Roach.P, Stibbard, R.Osborne, J Arnfield,S Selter, Transcription of prosodic and paralinguistic factors of emotional speech, Journal of International Phonetic Association. 28, 83-94.

[4] Crystal and Quirk, System of prosodic and paralinguistic features of English, 1964.

[5] J. Van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg: Progress in Speech Synthesis. Springer Publication 1995

[6] CMU: Festival FESTVOX synthesizer. www.festvox.org.

[7] CMU-Cambridge. Statistical language modeling Toolkit http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html

[8] AT&T Natural Voices ™ Synthesizer http://naturalvoices.att.com/