# Prediction and Realisation of Conversational Characteristics by Utilising Spontaneous Speech for Unit Selection

*Sebastian Andersson*[1], *Kallirroi Georgila*[2], *David Traum*[2], *Matthew Aylett*[13], *Robert A.J. Clark*[1]

[1]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[2]Institute for Creative Technologies, University of Southern California, Los Angeles, USA
[3]CereProc Ltd, Edinburgh, UK

J.S.Andersson@sms.ed.ac.uk kgeorgila@ict.usc.edu

## Abstract

Unit selection speech synthesis has reached high levels of naturalness and intelligibility for neutral read aloud speech. However, synthetic speech generated using neutral read aloud data lacks all the attitude, intention and spontaneity associated with everyday conversations. Unit selection is heavily data dependent and thus in order to simulate human conversational speech, or create synthetic voices for believable virtual characters, we need to utilise speech data with examples of how people talk rather than how people read. In this paper we included carefully selected utterances from spontaneous conversational speech in a unit selection voice. Using this voice and by automatically predicting type and placement of lexical fillers and filled pauses we can synthesise utterances with conversational characteristics. A perceptual listening test showed that it is possible to make synthetic speech sound more conversational without degrading naturalness.

**Index Terms**: speech synthesis, unit selection, conversation, spontaneous speech, lexical fillers, filled pauses

## 1. Introduction

Unit selection speech synthesis simulates neutral read aloud speech quite well, both in terms of naturalness and intelligibility [1]. For many applications such a neutral style is sufficient. However, there are other emerging applications where speech synthesis is expected to play a very important role, e.g. virtual human dialogue systems [2], speech-to-speech translation [3], etc. These applications require the speech synthesiser to be able to synthesise speech that gives an impression of the attitude, intention and spontaneity associated with everyday conversations. In other words, we need to simulate how people *talk* instead of how people *read* [4].

Spontaneous conversational speech exhibits many characteristics that are avoided or poorly modelled in current speech synthesis: pronunciation variation (elision, reduction) [5], disfluencies (word fragments, mispronunciations, hesitations, repetitions, repairs) [6, 7], voice quality and amplitude variations (attitude, emotions) [8], and paralinguistics (laughter, sighs, breathing) [9, 10, 11, 12]. We believe that by utilising spontaneous conversational speech that contains natural prosodic realisations of the above phenomena we can begin to build speech synthesisers with conversational characteristics.

The most obvious starting point for conversational synthesis is those aspects that are problematic for traditional speech synthesisers, namely, interjections (e.g. "wow", "ugh", "phew"), back-channels (e.g. "uh-huh"), lexical fillers (e.g. "so", "well", "you know"), and filled pauses (e.g. "uh", "um").

In this paper we focused on synthesising lexical fillers and filled pauses.[1] First we automatically predicted which fillers or sequences of fillers to insert in a sentence as well as where to insert them. Then we synthesised the sentence with a unit selection voice containing spontaneous speech. The results showed that we can make synthetic speech sound more conversational without degrading naturalness.

The structure of the paper is as follows: In section 2 we compare our approach with related work in spontaneous/conversational speech synthesis. In section 3 we describe our methodology. In section 4 we present our evaluation and provide results. Finally in section 5 we present our conclusion and propose future work.

## 2. Comparison with Related Work

Most research in speech synthesis has focused on neutral read aloud speech. But some attempts have been made to move towards synthetic speech with more spontaneous or conversational characteristics.

A more colloquial or conversational speaking style of synthetic speech was achieved in [5] by modelling sequences of pronunciation variants.

Prediction [10, 4, 13] and synthesis of filled pauses and hesitation have been approached with limited domain synthesis with spontaneous speech [9] and recordings of acted prompts [14, 12]. Whereas [14] developed a prosodic model of the hesitation before a filled pause and avoided coarticulation problems due to data sparsity by always synthesising filled pauses surrounded by silent pauses, [12] designed sentences to cover all word endings in French to synthesise the transition from speech into hesitation and laughter.

In [15] an alternative approach was taken to synthesise spontaneous conversational speech. They recorded a corpus of spontaneous speech from a female speaker in her everyday situations during 5 years. A concatenative voice was built, but concatenation was only allowed at phrase boundaries. Instead the focus was on synthesising utterances appropriate in conversational interaction.

Our approach builds on the above previous work with some modifications and extensions.

In terms of predicting where to insert fillers, we follow the paradigm of [4] and use a machine learning algorithm trained on conversational data. However, there are some differences between our approach and the method of [4]. First, we are not

[1]From now on and for the sake of conciseness we will use the term *fillers* to refer to both lexical fillers and filled pauses.

restricted to predicting only filled pauses but we predict a much wider variety of fillers or combinations of fillers. Second, we do not always realise fillers between silent pauses. Third, whereas [4] combine n-grams with a decision tree, our method is based on n-grams and the Viterbi algorithm. Our approach also differs from the method of [10] who predict filled pauses and breathing instances using finite state acceptor networks and heuristics respectively. Their models are trained on a corpus of spontaneous lecture monologues and use information from the words or phrases that precede the filled pauses. Instead we take into account the context both preceding and following fillers.

With regard to synthesising sentences with predicted fillers, we utilised spontaneous speech for synthesis in a similar manner to [9, 15], but unlike [12, 14] we preserved the natural prosodic realisations of filled pauses and associated hesitation in our data. But whereas [9] used limited domain synthesis and [15] only phrase level concatenation, we used approximately 2000 spontaneous utterances in diphone unit selection and blended them with read speech data to approach general speech synthesis with a conversational character.

# 3. Utilising Spontaneous Speech for Synthesis

## 3.1. Spontaneous Speech: Data and Processing

We recorded approximately 7 hours of unconstrained conversation between the first author and an American male voice talent. We had previously recorded the voice talent reading aloud sentences for synthesis (see section 3.2.1) and the conversation recordings were made in the same studio and with the same microphone.

The conversation was manually transcribed into words and aligned to the speech at the utterance level. Utterances with word fragments, overly reduced pronunciation or laughter were excluded from the voice build, and so were utterances where the voice talent "put on" different voices to display a third person, such as his wife or friends. The remaining utterances were still rich in spontaneous speech phenomena such as repetitions, lexical fillers and filled pauses, e.g.:

*"yeah it's it's a significant amount of swelling um [pause] more than like i'd say a bruise"*

One of the most noticeable differences between the spontaneous and read aloud data (recorded with the voice talent) was the speech rate. Thus to facilitate blending of spontaneous and read aloud material in synthetic utterances the spontaneous speech was slowed down 10% with SoundStretch [16]. The selected spontaneous utterances were forced aligned on phonemic level with the HTK toolkit [17]. To get a good alignment of the spontaneous speech we used acoustic models trained on the read aloud material.

In total 2120 spontaneous utterances with 83 min of phonetic material (excl. silent pauses) were included in the voice build. But more than 700 of these utterances consist only of back-channels (e.g. "yeah", "yeah yeah yeah" or "okay"), that although important for interaction do not add value in terms of phonetic coverage for synthesising out of database words or phrases.

## 3.2. CereVoice Speech Synthesis System

We used the CereVoice speech synthesiser developed by CereProc Ltd. This is a diphone unit selection speech synthesis engine available for academic and commercial use [18]. An input text sentence is converted into a sequence of phonemes and the Viterbi algorithm is used to search a database of speech to find an optimal sequence of diphone sized units to concatenate based on a sum of heuristically weighted target (linguistic) and join (acoustic) costs.

### *3.2.1. Speech Genres*

The core speech data resource of the two voices in this paper came from 2679 carefully and neutrally read aloud sentences to provide phonetic coverage of diphones in different syllable, word and phrase positions, which in total gave 104 min of phonetic material (excl. silent pauses). The voice built from only this material is henceforth referred to as the *read* voice.

In addition to the coverage material, CereVoice offers the capability of marking up the speech data with a genre tag to enable synthesis with different speaking styles with the same voice. At synthesis time a desired genre or speaking style can be requested by XML-tags. Units from undesired genres are then discarded before the Viterbi search if there are enough (50) units available from the desired genre. If enough units from the desired genre are not available, units from other genres are included in the Viterbi search [18].

In this paper we added the spontaneous speech data described in section 3.1 as a genre to enable speech synthesis of a more conversational style. The voice with the added spontaneous genre is henceforth referred to as the *spontaneous* voice.

## 3.3. Genre Switching Synthesis

Unit selection is more challenging with spontaneous speech than with carefully and consistently read aloud speech because of the greater acoustic variation between units. A genre switching approach was therefore adopted to find the best trade-off between selecting spontaneous units to achieve conversational quality, and selecting read aloud units to synthesise words that were not in the spontaneous speech database.

For a given input text sentence if a two word sequence in the input sentence existed in the spontaneous speech database then selection was biased towards spontaneous units, otherwise it was unbiased. This did not guarantee that units were selected from this sequence in the spontaneous speech database, but it did guarantee that we had suitable candidate units for the given word sequence.

## 3.4. Filler Prediction

Automatically predicting the type and placement of disfluencies (in our case fillers) is not trivial. Interestingly, although the problem of automatically detecting speech disfluencies has received much attention [6, 7], the same is not true for the automatic generation of disfluencies. The only other approaches we are aware of are [4, 10] mentioned above. And yet, incorrect predictions can dramatically decrease the naturalness of the synthesised utterances. This is because not every filler insertion is valid, e.g. consider the examples *"they so took um it away"* (invalid insertion) vs. *"so they took it um away"* (valid insertion).

During training we use the 2120 spontaneous utterances utilised in the spontaneous voice build, and generate word n-grams (3-grams and 2-grams). Thus, in the 3-gram case, given history $w_{i-2}w_{i-1}$ we calculate the probability of the next word $w_i$. Likewise for 2-grams.

During testing (on a held-out data set) the algorithm receives a sentence as input and must decide where to insert fillers

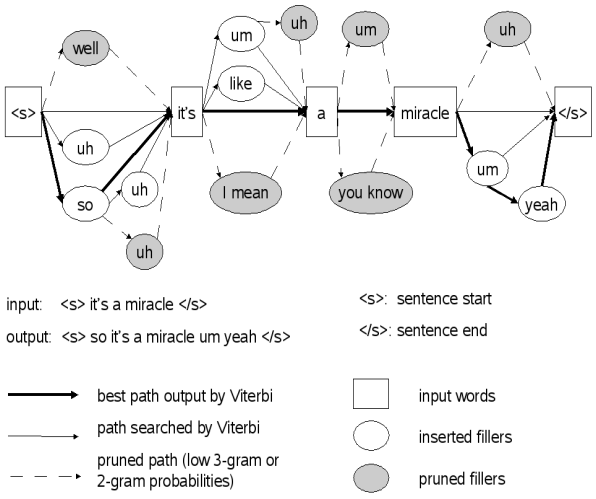


Figure 1: Example of filler prediction.

and which fillers to insert. The algorithm works as follows: Consider the word sequence $w_1$,$w_2$,...,$w_N$. For each word $w_i$ ($i$=1,...,$N$) we select, from our list of 3-grams, the 3-grams $w_{i-1}w_iw$ where $w$ is a filler. From these 3-grams we pick the ones with a probability over an empirically set threshold. Let $L$ be the number of these most probable 3-grams. Thus we have $L$ candidate insertions after $w_i$: $w_{i1}$,$w_{i2}$,...,$w_{iL}$. Then we can choose between two modes, *conservative* and *non-conservative*. In the conservative mode we compute the probabilities of the $L$ 3-grams $w_iw_{ij}w_{i+1}$ where ($j$=1,...,$L$), i.e. we look both backwards and forward to see if by inserting the filler $w_{ij}$ between $w_i$ and $w_{i+1}$ we will end up with something meaningful. In the non-conservative mode we compute the probabilities of the $L$ 2-grams $w_{ij}w_{i+1}$ where ($j$=1,...,$L$), i.e. we look only forward to see if inserting the filler before $w_{i+1}$ makes sense. Obviously since in the non-conservative approach we do not consider the left context it is very likely that we will allow more combinations (hence it is called non-conservative mode). After operating either on conservative or non-conservative mode we select the $M$ insertions with the highest probabilities (we use $M$=3). So now for each word $w_i$ we have $M$ candidate fillers that can follow.

We repeat the process of the previous paragraph to generate sequences of fillers. Now the input sentence is the sentence with the previously inserted fillers. For our experiment we allowed sequences of fillers of length 2.

The next step is to use Viterbi and select the combination of insertions along the whole sentence that will lead to the highest overall probability. Obviously we do not want to insert fillers after every word in the sentence because that would sound awkward. Thus we can set a threshold $T$ for the maximum number of fillers allowed (we used $T$=5) and Viterbi will take this threshold into account. We can also adjust the algorithm either to generate always the best sequence or the n-best sequences and then choose randomly among them. This is to ensure some variation in the output.

For our experiment we used both the conservative and the non-conservative modes and the sequence with the highest probability. Before realising the sentence with the speech synthesiser, silent pauses are added after a predicted filler (or sequence of fillers) to reinforce the grouping of the hesitation (normally preceding a filler) and the filler, and to avoid concatenation artefacts due to data sparsity. The algorithm can generate 17 different types of lexical fillers and 2 different types of filled pauses ("uh", "um") as well as their combinations. An example is given in Figure 1.

# 4. Evaluation

A perceptual listening test was designed to evaluate if we could utilise spontaneous speech data to synthesise speech with a more conversational character than synthesis based only on read aloud speech.

The test sentences were randomly selected from a held-out set of the transcribed conversation. Original disfluencies and fillers were removed from the held-out transcripts (using the algorithm of [7]) and replaced with predicted fillers or sequences of fillers. To better evaluate the potential of predicting and synthesising a wide variety of types and placements of fillers, we restricted the selection of test sentences so that we have the same filler sequence in at most two sentences. We selected 15 sentence pairs to synthesise:

- Sentences with no fillers, e.g.:
  *"it's a different character for me"*
- Sentences with predicted fillers, e.g.:
  *"so it's um [pause] a different character for me"*

Both the sentences with and without predicted fillers were synthesised with the *read* voice. In the listening test they were compared to the sentences with predicted fillers synthesised with the *spontaneous* voice giving two test conditions:

I) Sentences with predicted fillers synthesised with the *spontaneous* voice vs. sentences with predicted fillers synthesised with the *read* voice.

II) Sentences with predicted fillers synthesised with the *spontaneous* voice vs. sentences without fillers synthesised with the *read* voice.

## 4.1. Listening Test

A web-based listening test was carried out with 30 volunteering participants. The 15 sentence pairs for the two conditions were played to the participants in randomised order and also mirrored. In total each participant listened to 60 sentence pairs of synthetic speech and were asked about their opinions on two different aspects:

- *Which utterance in the pair sounds more like in an everyday conversation (as opposed to e.g. someone reading from a script)?*
- *Which utterance in the pair sounds more natural (regardless if it sounds conversational or not)?*

The participants could express preference for either utterance in the pair ("A" or "B") or select a no-preference option ("Equal").

## 4.2. Results

The perceptual judgements have been collapsed over all utterances and are shown for both comparisons as percentages in Figure 2.

The results were calculated with the binomial test with two sided 95% confidence interval. The number of times that the participants judged the quality as "Equal" (EQ) was split in half

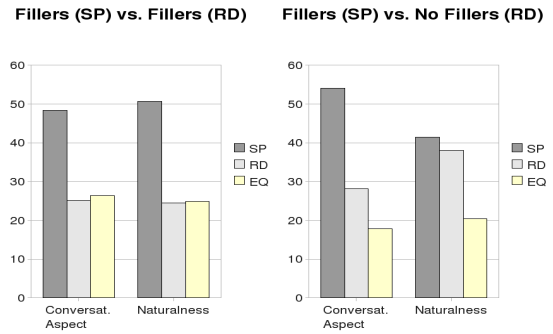


Figure 2: Percentage (%) of perceptual judgements for "Conversational" and "Natural" quality of synthetic speech when comparing the *spontaneous* (SP) voice with fillers to the *read* (RD) voice with and without fillers respectively. "Equal" (EQ) expressed no preference.

and assigned to the judgements for the *spontaneous* (SP) and *read* (RD) voice respectively. The null hypothesis that there was no preference between the voices was then tested.

The *spontaneous* voice with fillers was perceived as significantly more conversational than the *read* voice with fillers (62% for *spontaneous*, 38% for *read*, $p = 4.2 \times 10^{-12}$). The *spontaneous* voice with fillers was also perceived as significantly more natural than the *read* voice with fillers (63% for *spontaneous*, 37% for *read*, $p = 3.3 \times 10^{-15}$). This means that it is not sufficient to just insert fillers in text, but it is essential to have appropriate realisations of fillers in the voice, otherwise naturalness is negatively affected.

The *spontaneous* voice with fillers was perceived as significantly more conversational than the *read* voice without fillers (63% for *spontaneous*, 37% for *read*, $p = 5.7 \times 10^{-15}$). The *spontaneous* voice with fillers and the *read* voice without fillers were not perceived as significantly different in terms of how natural they sounded (52% for *spontaneous*, 48% for *read*, $p = 0.33$). This means that we can include spontaneous speech in synthesis to achieve a more conversational style without decreasing the general naturalness.

## 5. Conclusion and Future Work

We showed that by utilising spontaneous speech and predicting type and placement of lexical fillers and filled pauses, we were able to synthesise speech with a more conversational character, with on average no loss of naturalness, compared to synthetic speech based only on read aloud data.

In general the test sentences did not contain many concatenation errors or "bad joins". When the quality was not good it seemed to be more an issue of inappropriate selections, e.g. overly reduced pronunciation in the wrong places, and that blending of spontaneous and read aloud speech needs to be investigated more thoroughly.

Training filler predictions and building a synthetic voice from the same speech data meant that we could limit the insertions of spontaneous speech phenomena to those that we could confidently synthesise. The fact that we have a substantial amount of spontaneous speech data that we so far have not utilised means that we could extend our predictions to, for example, simple disfluencies (e.g. function word repetitions) and paralinguistics (e.g. laughter, sighs, throat clearings) using the same approach.

## 6. Acknowledgements

The first author is supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568). This work was sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). The content does not necessarily reflect the position or the policy of the U.S. Government, and no official endorsement should be inferred.

## 7. References

[1] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The Blizzard challenge 2008," in *The Blizzard Challenge*, Brisbane, Australia, 2008.

[2] D. Traum, W. Swartout, J. Gratch, and S. Marsella, "A virtual human dialogue model for non-team interaction," in *Recent Trends in Discourse and Dialogue*, L. Dybkjaer and W. Minker, Eds. Antwerp, Belgium: Springer, 2008, pp. 45–67.

[3] T. Schultz, A. Black, S. Vogel, and M. Woszczyna, "Flexible speech translation systems," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 403–411, 2006.

[4] J. Adell, A. Bonafonte, and D. Escudero, "Filled pauses in speech synthesis: Towards conversational speech," in *10th International Conference TSD 2007*, Pilsen, Czech Republic, 2007, pp. 358–365.

[5] S. Werner and R. Hoffmann, "Spontaneous speech synthesis by pronunciation variant selection – a comparison to natural speech," in *Interspeech*, Antwerp, Belgium, 2007, pp. 1781–1784.

[6] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[7] K. Georgila, "Using integer linear programming for detecting speech disfluencies," in *NAACL-HLT 2009: Short Papers*, Boulder, Colorado, USA, 2009, pp. 109–112.

[8] E. Krahmer and M. Swerts, "On the role of acting skills for the collection of simulated emotional speech," in *Interspeech*, Brisbane, Australia, 2008, pp. 261–264.

[9] S. Sundaram and S. Narayanan, "Spoken language synthesis: Experiments in synthesis of spontaneous dialogues," in *IEEE Speech Synthesis Workshop*, Santa Monica, California, USA, 2002.

[10] ——, "An empirical text transformation method for spontaneous speech synthesizers," in *Eurospeech*, Geneva, Switzerland, 2003, pp. 1221–1224.

[11] N. Campbell, "Conversational speech synthesis and the need for some laughter," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1171–1178, 2006.

[12] D. Cadic and L. Segalen, "Paralinguistic elements in speech synthesis," in *Interspeech*, Brisbane, Australia, 2008.

[13] J. Adell, A. Bonafonte, and D. Escudero, "Statistical analysis of filled pauses' rhythm for disfluent speech," in *SSW6*, Bonn, Germany, 2007, pp. 223–227.

[14] ——, "Disfluent speech analysis and synthesis: A preliminary approach," in *3rd International Conference on Speech Prosody*, Dresden, Germany, 2006.

[15] N. Campbell, "Towards conversational speech synthesis; lessons learned from the expressive speech processing project," in *SSW6*, Bonn, Germany, 2007, pp. 22–27.

[16] http://www.surina.net/soundtouch/soundstretch.html, Oct. 2009.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2006.

[18] J. Andersson, L. Badino, O. Watts, and M. Aylett, "The CSTR/CereProc Blizzard entry 2008: The inconvenient data," in *The Blizzard Challenge*, Brisbane, Australia, 2008.