interspeech 2008 incorporating SST 08

# Synthesis by Generation and Concatenation of Multiform Segments

*Vincent Pollet, Andrew Breen*

Text-To-Speech Research, Nuance Communications, Merelbeke, Belgium

vincent.pollet@nuance.com, andrew.breen@nuance.Com

## Abstract

Machine generated speech can be produced in different ways however there are two basic methods for synthesizing speech in widespread use. One method generates speech from models, while the other method concatenates pre-stored speech segments. This paper presents a speech synthesis technique where these two basic synthesis methods are combined in a statistical framework. Synthetic speech is constructed by generation and concatenation of so-called "multiform segments". Multiform segments are different speech signal representations; synthesis models, templates and synthesis models augmented with template information. An evaluation of the multiform segment synthesis technique shows improvements over traditional concatenative methods of synthesis.

**Index Terms**: speech synthesis, statistical selection, HMM, multiform segments

## 1. Introduction

Model-based and concatenative-based text-to-speech (TTS) systems differ in many ways. In terms of principle, concatenative-based [1] systems select basic units in the form of waveform segments from a speech corpus and stitch these together with a minimum of signal manipulation while model-based systems rely heavy on signal manipulation as speech waveforms are decomposed into speech parameters, modeled and reconstructed.

Concatenative systems produce variable quality synthetic speech, which at best is highly natural. This variation in quality is dependent on the length of continuous speech selected from a unit inventory. Limited domain concatenative systems, which tend to return long stretches of stored speech, produce very natural synthesis. In contrast, model-based systems have more consistent speech quality but with a synthetic "processed" character. The generated speech is smooth and stable, as the system shows good predictable behavior with respect to unseen contexts.

In terms of footprint, concatenative systems employ large unit inventories as the speech quality is critically dependant on the size and accuracy of the unit inventory. In addition, simple speech coding of the unit inventory leads to a large system footprint. Model-based systems on other hand, work well on small unit inventories [2]. The heavy (statistical) modeling of speech parameters rules out odd errors (such as; labeling and segmentation errors) in the unit inventory and results in a small system footprint.

The differences between the systems have important practical implications. The construction of a new text-to-speech voice for commercially viable concatenative systems is costly in terms of time and labor: a large corpus of speech needs to be recorded, accurately labeled and segmented into basic speech units [3]. Constructing a new voice for a model-based system is less costly as less data and consequentially less effort is needed. The synthesis models may also be transformed in order to derive multiple voices from a single donor voice using only a very small amount of new observation data [4].

In this paper, we present the **Multiform Segment** (MFS) synthesis technique, the aim of which is to combine the benefits of the model-based and the concatenative approaches. **Statistical and probabilistic** techniques are adopted from the model-based methods to form the basic framework of the system. As a consequence this enables automatic processing, consistency, auto-tuning and good generalization characteristics that are common to model-based methods. Generalization is achieved by the dynamic construction of unseen sounds. Both natural speech fragments and models are used in the speech construction process. The natural quality which can be achieved by selecting and concatenating natural speech waveform fragments is adopted from the concatenative method.

MFS synthesis also employs characteristics of **speech perception** during the speech construction process. It takes advantage of the fact that there is perceptual invariance in speech [5]. Humans do not perceive and need high detail at all times. For machine-based speech generation this motivates the question: "can some detail in speech production be sacrificed some of the time?" Reformulating this question in context of MFS synthesis yields: "can certain speech segments be abstracted, interchanged and re-used without causing degradation during speech generation?" The MFS algorithm answers this question and employs this knowledge to **sequence** multiform segments. As a consequence of decisions which segments can be abstracted, the MFS algorithm is able to determine which segments within a sequence play a perceptual important role.

This paper is organized as follows. In section 2, we briefly introduce and define the basic principles of MFS synthesis. In section 3, we present the statistical model segments which form the basis of our unified statistical framework. In section 4, we show how to select template segments from a unit inventory given the model segments. In section 5, we address how to combine and sequence model segments and template segments. In section 6, results from experiments to evaluate the performance of a MFS synthesis against traditional concatenative systems are given, and finally in section 7, conclusions are drawn from this work.

## 2. Multiform Segment synthesis

MFS synthesis is defined as a method of employing segments having a plurality of different speech representational structures for speech construction. A segment or a sequence of segments may be:

- A **template segment** $O_T$. This is an instance of speech segment, for example the speech waveform corresponding with the last two phonemes \at\ of the word 'cat'. This template may be represented as simply as a phonetic and prosodically labeled fragment of recorded speech or alternatively it may

be represented in other ways, for example, compressed by means of a speech coder.

- A **model segment** $O_M$ is an abstraction of a speech segment, such as a statistical-mathematical model produced by a hidden Markov model (HMM) training process on an inventory that would contain multiple exemplars of the phonemes *at*\. The abstraction may even be at different streams or layers, for example, only the harmonic components of a source-filter speech representation may be abstracted.

The aim of the MFS algorithm is to determine the best sequence of segments $O^*$ for a given input text, where the segments consist of the best combination of model segments and template segments:

$$O^* = \bigcap_{i=1}^{N} \left\{ O_i \middle| O_i = {O_{M_i}}^* \vee {O_{T_i}}^* \right\} \quad (1)$$

# 3. Model Segments

In this paper, HMM modeling for speech synthesis is used to create model segments [2] however in principle any other synthesis model can be used. A state-tree search is performed to find the best model segments. The trees (for each state-component) are constructed as part of the state-clustering process performed during HMM training. In HMM training for speech synthesis, extreme speaker dependant training is performed. Continuous density HMMs are built on high-precision acoustic observations. Each HMM is defined by a phonetic-prosodic context vector.

An important aspect in HMM synthesis is the construction of the acoustic parameter trajectories from the HMMs. In this algorithm, information on static and dynamic state output distributions, is used in order to generate smooth and natural varying acoustic parameter trajectories [6]. This algorithm can be improved if additional observations on parameter trajectories are added. In this work, the variance of acoustic parameter trajectories within phrases of about the same length (i.e. a similar number of voiced and unvoiced states) and maybe more importantly, the variance of acoustic parameters of natural speech or templates of fragments within the same phrase, are used as additional observations. This information can aid the trajectory generation algorithm as it provides additional information upon the dynamics of natural speech instances over longer stretches. The formulation of trajectory generation algorithm can consequently be reformulated as following:

$$L = \log\left\{ P\left(O_M \middle| Q, \lambda\right)^{\varpi} P\left(v(C) \middle| \lambda_{vg}\right)^{\frac{\eta}{100}} P\left(v(C) \middle| \lambda_{vt}\right)^{1-\frac{\eta}{100}} \right\} \quad (2)$$

The first term is the probability of finding a model observation given the state-mixture sequence $Q$ and continuous density HMM $\lambda$. This is the same term as presented in [6]. The second term is the probability of the variance of the acoustic parameters $C$ given a crude phrase model. The third term expresses the probability of variance of acoustic parameters within some fragments of the same phrase. The solution is obtained by finding $C$ that maximizes $L$. As all probability terms are a function of $C$, a solution can only be found using an iterative function of $C$. The Newton-Raphson gradient method is adopted and results in a fast convergence. The first term is scaled by the dynamic dimension coefficient $w$. The second and third terms are scaled by $\eta$, which we define as the model-template ratio (MTR). The MTR is the percentage of model segments used for constructing a message containing $N$ segments. MTR can be expressed with or without segment duration normalization. As a result of MTR scaling, when more template segments are used to sequence multiform segments, more influence on the variance of template segments is put during the parameter trajectory generation. In the next section, we address how to obtain the variance of acoustic parameters of regions of natural speech within the same sentence.

# 4. Template Segments

## 4.1. Cost functions

In traditional concatenative TTS systems, a number of metrics are used to define target and join costs. The target cost estimates how features of a recorded unit (or template) match a specified feature vector. The join cost attempts to express how good units will be perceived when joined together. The optimal sequence of template segments is found by minimizing both costs. This is typically achieved by a Viterbi search. These traditional cost computations employ a great deal of heuristics. For example, one can compute a cepstral or another form of acoustical distance between two units, however the question remains, how does this distance relate to observations of natural adjacent speech units? As an illustration, the cepstral distance between a vowel and an obstruent can be large even though the context may be perfectly natural. A large distance would mean a significant cost contribution, which would adversely affect the selection of units such that a less optimal and perceptual salient join would be chosen over the natural join. Concatenative TTS systems typically circumvent this problem by applying some form of manually tuned weight scaling triggered by a number of events such as phonologic and prosodic context.

## 4.2. Probabilistic formulation

Similar to the proposed approach in [7], heuristics are replaced by statistics. In this work, model segments are used as basis of the probabilistic formulation.

$${O_T}^* = \arg\max_{O_T} P\left(O_T \middle| F\right) \quad (3)$$

$$F_i = \left\{ f_i(1), .., f_i(k) \right\} \quad (4)$$

In order to obtain the best sequence of template segments $O_T = o_{Ti}, .., o_{TN}$, the probability of candidate template segments given a collection of target features $F$ must be maximized. The target feature collection $F$ includes symbolic and acoustic target features $f$ for each *i-th* template segment. In a probabilistic formulation, the probability of finding a template segment is assumed to be dependent on the previous template. In this formulation, all the acoustics are modeled by a continuous density HMM $\lambda$ with state-mixture sequence $Q$ and each HMM is annotated by its full phonetic-prosodic vector. In analogy with the first term of (2), the probability of a template given a collection of target features then becomes the joint probability of a template segment given the previous template and the state-mixture sequence $Q$ and continuous density HMM $\lambda$.

$$P(O_T|F) = \prod_{i=1}^{N} P(O_{T,i}|O_{T,i-1}, Q, \lambda, i) \quad (5)$$

This can be rewritten as the probability of a template segment $i$, given the HMM $\lambda$ and state sequence $Q$ multiplied with the probability of a left boundary feature $l$, after having a $r$, right boundary feature of the previous template observation from the feature collection $F$.

$$P(O_{T,i}|O_{T,i-1}, Q, \lambda, i) = P(O_{T,i}|Q, \lambda) P(l(O_{T,i})|r(O_{T,i-1}), F, i) \quad (6)$$

In analogy to target and join costs, a target and a join probability can now be introduced. The target probability (7) is the joint probability of the features from feature collection $F$ of a template segment given the feature parameter state output distributions of the HMM $\lambda$. This term is straightforward to compute. The join probability is the second term of (6) and is the probability of the left boundary feature of the current template observation given the right boundary feature of the previous template and this for feature collection $F$. This term can be computed using a language modeling technique [8] which uses n-grams to express the probability that one template segment frame follows another template segment frame.

$$P(O_{T,i}|Q, \lambda) = \prod_{j=1}^{k} P(f_i(j, O_{T,i})|\lambda_{f_i(j)}) \quad (7)$$

The optimal sequences of template segments is computed by maximizing both target and join probabilities, which can be done by means of a Viterbi search.

## 5. Sequencing multiform segments

The best sequence of multiform segments is obtained by maximizing the probability of segments given that a segment is either a model segment or a template segment.

$$O^* = \arg\max_O P(O|O = O_M{}^*, O_T{}^*) \quad (8)$$

Where,

$$P(O|O = O_M{}^*, O_T{}^*) = \prod_{i=1}^{N} P(O_i|O_i = O_{M,i}{}^* \vee O_{T,i}{}^*, i) \quad (9)$$

The probability of a multiform segment being a model segment depends on three categories of cues. These are:

1. **Phonologic cues**; the phonetic context $Sc_M$ and position $Sp_M$ of a model segment.
2. **Acoustic cues**; the duration $d_M$ and unvoiced distribution $\lambda_{M,Uv}$ of a model segment.
3. **Channel cues**; the regeneration capability $G$ of a model segment.

$$P(O_i|O_{M,i}{}^*) = P(O_i|Sc_{M,i}, Sp_{M,i}, d_{M,i}, \lambda_{M,Uv,i}, G) \quad (10)$$

These cues form the foundation of our speech perception model as they enclose perceptual characteristics of speech. Phonologic cues are, for example, that (stressed) vowels and syllable nuclei are perceptual important. At such locations, the selection of template segments is favored over the selection of model segments. In contrast, as humans are perceptually less sensitive to the fine structure of obstruents and nasals, these types of sound can be abstracted. For example, model segments can be used for syllable-initial nasals. High frequent mono-syllabic words are exceptions to this as they often contain pronunciation variants and phoneme reductions and are therefore not split up if they can be fetched as a unity from a template inventory.

An example of a phonologic-acoustic cue is given by the fact that obstruents generally join well with other phonetic classes and so can be modeled. An example of an acoustic cue is: for consonants, duration seems to be more important than spectral detail. Here also model segments may be considered.

The channel cues are defined by the limitations of the parametric synthesizer model. The regeneration quality can be regarded as a function of the acoustic parameters. Certain speakers are more sensitive to modeling for parametric synthesis than others. Where too much degradation (i.e. buzzyness) is introduced the usage of model segments is discouraged. Also we experienced that an increase in unvoiced energy for certain model segments helps the multiform segment sequencing.

The probability of a multiform segment being a template segment is:

$$P(O_i|O_{T,i}{}^*) = (1 - P(O_{M,i}{}^*)) P(O_{T,i}{}^*|\lambda_i) \quad (11)$$

Where $P(O_{T,i}{}^*|\lambda_i)$ is the probability of a template segment given the continuous density HMM $\lambda$. This term can be computed as a side-result of a forced alignment algorithm and corresponds roughly to the template likelihood score [9].

A scheme of the MFS TTS system is depicted in figure 1 below.

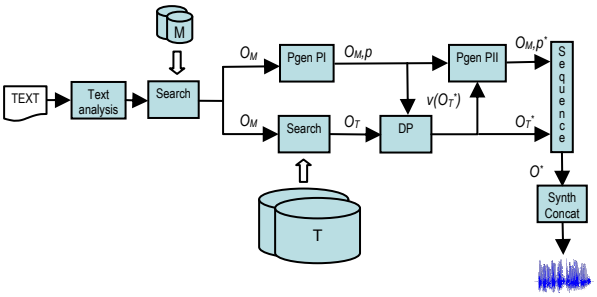


Figure 1: *Scheme of a MFS TTS system.*

Text is fed into the text-analysis process. Next, a search is performed by matching the phonetic and prosodic context vectors to the HMMs stored in the model inventory. As result, a sequence of model segments $O_M$ is obtained. The model segments are used to direct a search of template segment candidates $O_T$ in the template inventory. In the parallel branch, a first parameter trajectory generation pass (*Pgen PI*) operates on the model segments and generates the acoustic parameter trajectories $p$ given the model segments. The model segments with the parameter trajectories $p$ and template segment candidates are input into a dynamic programming

algorithm (*DP*). As a result, the best template segment sequence $O_T^*$ is obtained. The variance of acoustic parameter trajectories of the best template sequences is fed into a second parameter generation algorithm (*Pgen PII*) which regenerates the acoustic parameters *p*. Speech parameter trajectories $O_{M,p}^*$ are obtained of which the variance reassembles the variance of the best template segments. This is done to combine seamlessly template segments with model segments. Next, the best models and template segments are sequenced. This sequence $O^*$ of best multiform segments is send to the synthesizer-concatenator. The acoustic parameter trajectories are converted to speech waveforms and concatenated with the template segments, which yields the speech output-waveform.

# 6. Evaluation

## 6.1. Setup

A formal listening experiment similar to [10] was organized in-house using an experimental MFS TTS system. From the 80 participants, the large majority were speech professionals. The MFS TTS system (systems: R,Q,P and O) was tested against other laboratory concatenative systems (systems: C and D). Natural speech (N) was also included in the test set. As model segments and statistical framework for the MFS system, continuous density HMMs consisting of five states left-to-right with no skips using single Gaussian distributions as output probabilities with a diagonal covariance matrix were used. For the MFS system, different average model-template ratios without duration normalization were evaluated: ($\overline{MTR(R)} = 2, \overline{MTR(Q)} = 32, \overline{MTR(P)} = 7, \overline{MTR(O)} < 1$). The model-template ratio can be controlled in the sequencer (*Figure 1*). The same speech corpora were used for both systems. A medium size corpus was used to generate systems; C, Q and O and a small corpus containing only a few hundred phrases was use to generate systems D and R. 8 different text categories were used; news, conversation, numbers, dates, currency, isolated words, semantic unpredictable sentences, carriers with slots and sentence sets (or paragraphs). Both preference and MOS tests were held where different aspects of quality where evaluated such as; intelligibility, naturalness, non-monotonicity, fluidity, comprehension and pleasantness. Significance analysis ($p<0.05$) was performed using the Tuckey honest significant difference (HSD) test and ANOVA.

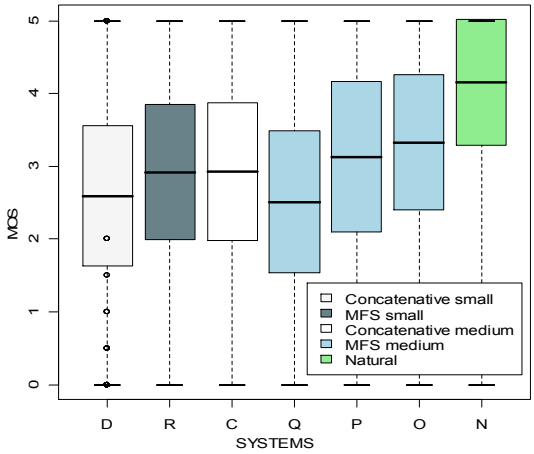


Figure 2: *MOS test results.*

## 6.2. Results

The results of the formal evaluation are depicted in figure 2. On the small size corpus, the MFS system *R* is found to be better than the concatenative system *D*. Also the small MFS system *R* is of equal quality to the concatenative system *C* which uses a medium size corpus. Differences in score between *R-C*, *D-Q* and *P-R* are not statistically significant. The MFS system with the highest average MTR Q did not perform well on the medium size corpus. At a more relaxed MTR setting *P* and *O*, the MFS system was better than the concatenative system *C*.

All preference tests showed a clear preference of the MFS system over the concatenative system. As general preference, measured over multiple test categories, 58% was obtained for the MFS system, against 23% for the concatenative system and 19% which had no preference.

Post-test analysis indicated that the less good performance of system *Q* was mainly caused by volume differences and concatenation artifacts. Since resolving the causes, more recent empirical experiments have shown that MTR levels up to 60 can yield good MFS synthesis.

# 7. Conclusions

A multiform synthesis technique was presented that shows improvements compared to a laboratory concatenative system. Next to the improvements in quality, the MFS system has appealing characteristics as a novel technique; it employs statistical and perceptual principles for the construction of speech.

# 8. References


[1] Coorman, G., Fackrell, J., Rutten, P. and Van Coile, B., "Segment selection in the L&H Realspeak laboratory TTS system", Proc. ICSLP, 2:395–398, 2000.

[2] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S., "Speech synthesis using HMMs with dynamic features", Proc. ICASSP, Atlanta, USA, pp. 389-392, 1996.

[3] Pollet, V. and Coorman, G., "Statistical corpus-based speech segmentation", Proc. ICSLP, Jeju Island, South Korea, pp. 1929-1932, 2004.

[4] Yamagishi, J., Masuko, T., and Kobayashi, T., 2004. "MLLR adaptation for hidden semi-Markov model based speech synthesis", Proc. ICSLP, Jeju Island, South Korea, pp. 1213-1216, 2004.

[5] Liberman, A.M., Harris, K.S., Hoffman, H.S., and Griffith, B.C., "The discrimination of speech sounds within and across phoneme boundaries", Journal of Experimental Psychology 54: 358-368, 1957.

[6] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. and Kitamura T., "Speech parameter generation algorithms for HMM-based speech synthesis", Proc. of ICASSP, pp.1315-1318, 2000.

[7] Allauzen, C., Mohri, H. and Riley M., "Statistical Modeling for Unit Selection in Speech Synthesis", Proc. ACL Barcelona, Spain, July 2004.

[8] Taylor, P., "Unifying Unit Selection and Hidden Markov Model Speech Synthesis", Proc. ICSLP, 2006.

[9] Amdal, I., Johnsen, M.H. and Svendsen T., "Log Likelihood Ratio Based Annotation Verification of a Norwegian Synthesis Database", Proc. Norsig 2006, Reykjavik, June 2006.

[10] Black, A.W. and Tokuda, K., "The Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets", Proc. ICSLP, 2005.