# Creating Expressive TTS Voices for Conversation Agent Applications

Andrew Breen

Nuance Communication, Norwich, United Kingdom
abreen@nuance.com

**Abstract.** Text-to-Speech has traditionally been viewed as a “black box” component, where standard “portfolio” voices are typically offered with a professional but “neutral” speaking style. For commercially important languages many different portfolio voices may be offered all with similar speaking styles. A customer wishing to use TTS will typically choose one of these voices. The only alternative is to opt for a “custom voice” solution. In this case, a customer pays for a TTS voice to be created using their preferred voice talent. Such an approach allows for some “tuning” of the scripts used to create the voice. Limited script elements may be added to provide better coverage of the customer’s expected domain and “gilded phrases” can be included to ensure that specific phrase fragments are spoken perfectly. However, even with such an approach the recording style is strictly controlled and standard scripts are augmented rather than redesigned from scratch. The “black box” approach to TTS allows for systems to be produced which satisfy the needs of a large number of customers, even if this means that solutions may be limited in the persona they present.

Recent advances in conversational agent applications have changed people’s expectations of how a computer voice should sound and interact. Suddenly, it’s much more important for the TTS system to present a persona which matches the goals of the application. Such systems demand a more flamboyant, upbeat and expressive voice. The “black box” approach is no longer sufficient; voices for high-end conversational agents are being explicitly “designed” to meet the needs of such applications. These voices are both expressive and light in tone, and a complete contrast to the more conservative voices available for traditional markets. This paper will describe how Nuance is addressing this new and challenging market.

**Keywords:** Expressive text-to-speech, voice talent selection, conversational style.

## 1 Introduction

The commercial importance of Text-to-Speech (TTS) systems has been steadily growing year on year, with systems being deployed in a wide variety of markets ranging from low-end embedded devices such as toys and cell phones, to in-car solutions for navigation, and finally deployed as large scale systems for Enterprise solutions

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 1–14, 2014.

used for directory assistance, customer care and most recently a host of novel domains such as news reading and information query. Each market has specific demands on the technology; on embedded devices TTS systems must compete for limited “real estate“, while large server based systems must be computationally efficient and able to service hundreds of simultaneous requests in real-time while providing high quality synthesis.

The success of TTS in these different markets has been due to a combination of factors, most notably the development and adoption of the right technology for a given market and an understanding of how to make the technology work effectively for commercial applications. TTS in the market place must be robust to a broad range of input material while offering an “acceptable” level of performance and quality.

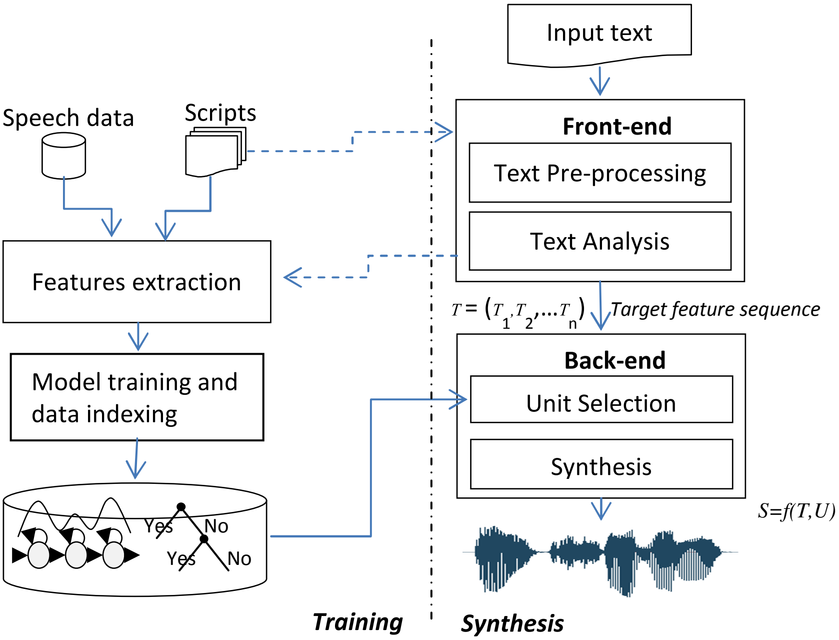

**Fig. 1.** Diagram of the training and synthesis phases in speech synthesis; during training a dbase of indexed units is created. During synthesis the search function $S=f(T,U)$ is used to obtain an sequence of units which optimally matches each target with units in the inventory.

Text-to-speech (TTS) systems have developed over years of research [1,2], resulting in a relatively standardized set of components as shown in Figure 1. The Front-end (FE), which derives information from an analysis of the text, and the Back-end (BE), which uses this information to search an indexed knowledge base of pre-analysed speech data. Indexed data most closely matching the information provided by the front-end is extracted and used by a speech synthesizer to generate synthetic speech. The pre-analysed data may be stored as encoded speech or as a set of parameters used to drive a model of speech production or as in hybrid systems a combination of both. It can be argued that recent commercial deployments of TTS have forced the pace of development in the back-end more than the front-end, although as this paper will discuss, this situation may now be changing. Back-end developments have consolidated into two broad categories; unit selection followed by waveform concatenation and unit selection followed by parametric synthesis, each approach having specific benefits. Waveform concatenation [4,5,6] currently offers the highest segmental speech quality but such systems are large and inflexible. Parametric synthesis systems [7] are robust to data compression, flexible and produce a more consistent quality, but currently suffer from a “synthetic” speech quality. At the moment, waveform concatenation methods are the most widely deployed solutions, parametric systems being limited to deployments which have strict computational and memory constraints.

This practical approach to development has lead to what some call the “encoding of ignorance” within modern commercial systems. Such systems have focused on the production of an overall solution, deploying methods which afford improvements in quality leading to great technology adoption, but do not attempt to offer significant insights into the underlying mechanisms of speech production. This pressure to feed the increasing demands of applications has resulted in a technological cul-de-sac, which is forcing researchers to re-evaluate well-established methods.

The paper is divided into 7 sections, each section describing in detail the steps taken by Nuance to address one aspect of this growth in demand; the creation of “characterful” synthesis systems for conversational agent applications. Section 2 will review an often overlooked but important element in successful system design: the selection of the voice talent. Section 3 provides an overview of the steps taken in creating an appropriate recording script used to build the synthesis voice. Section 4 discusses the importance of prosody in expressive voices and how it is used within these systems. Sections 5 and 6 describe the different synthesis methods investigated. Finally Section 7 provides results and conclusions.

## 2 The Voice

As previously stated in Section 1, commercial systems have focused on developing techniques which improve the adoption of synthesis. For a system to be deployed it must meet the acceptance criteria of a customer. This includes objective metrics such as pronunciation accuracy, but it also includes subjective metrics such as how pleasant the voice is and how well it matches the persona being designed within the whole application. Section 4 will discuss in detail the technical challenges facing TTS systems when asked to produce specific speaking styles. This section will focus on the interaction between the characteristics of the voice of the recording talent and the demands of a specific speaking style.

Traditional TTS applications have been dominated by basic information retrieval and confirmation applications. This is in part because of the demand for such services in the broader speech market, but also because the limitations of the technology have played well in these domains. Directory, banking and booking services have preferred personas which are mature, conservative and relatively slow paced. Voice selection and creation processes have been tuned over the years to cater for these markets, with large “portfolios“ of voices being developed in many languages. Where customers have requested a specific voice talent, such talents have be tutored to produce recordings in a specific style which works well with traditional TTS technologies. Portfolio voices are a combination of speaker, style and script and are designed to cater for the widest possible use case. The constraints imposed by technology in script design are considered in Section 3. “Custom voices“ are a useful supplement to the portfolio model. In such cases in addition to the choice of voice talent, systems may be tailored in script design to ensure that a particular customer‘s domain is well represented by the voice. The ultimate expression of this is “gilded phrases“, where specific recordings are stored and reproduced by the TTS system unadulterated. The application of gilded phrases in the development of expressive voices is discussed in Section 6.

The choices of speaker and style, as with many other commercial topics, are subject to changes in fashion. Recently a trend has emerged for more lively, youthful and dynamic personas that do not work well with the traditional methods of synthesis which have been heavily tailored to the pre-existing market. In order to better understand this relationship between speaker and speaking style a series of extensive MOS evaluations were conducted. Figure 2 shows the results of a MOS evaluation which compared a high quality portfolio voice, recorded in the standard portfolio style, with an example of a voice designed for the conversational agent market. 23 native US subjects were asked to score on clarity of pronunciation, naturalness, and overall impression. Scores were measured on a 5 point scale with 5 being the highest score. In order to evaluate the significance of different types of text material on subjective preference, two tests were conducted: one test composed of short length material e.g. navigation and prompt domain, and another using longer “passage“ length material e.g. news. The experiments suggest that there is a marked preference for the conversation style in the shorter material, and a slight preference for the traditional style in the longer material.

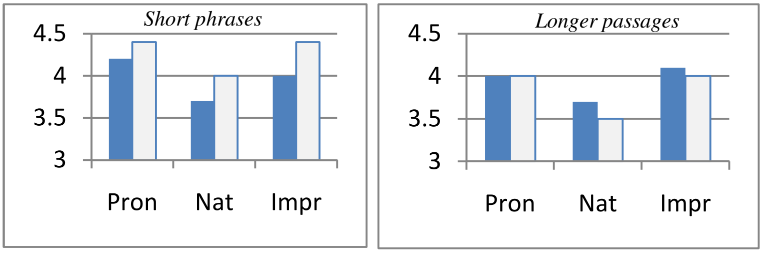

**Fig. 2.** Results from two MOS evaluations comparing different speakers and speaking styles on two types of text material: short phrases and longer passages. The results for the portfolio voice appear as “solid fill“ bars while the results for the conversational style are shown as “no fill“ bars. The results suggest that for shorter material in particular there is a strong preference for a more youthful speaker and dynamic style.

These experiments strongly suggested that in order to meet the demands of conversational agent applications, a more youthful voice is needed. A short list of 3 voices talents were selected from 25 candidates. An extensive MOS evaluation was then conducted to determine which of the shortlisted candidates met the requirements of pleasantness, dynamism and friendliness. As a baseline the same portfolio voice used in the previous experiment was included. Evaluations were conducted using 40 native US subjects. Subjects were asked to rate the voices on pleasantness and ease of listening. Each voice talent recorded 5 passages of 15-20 seconds. The material was randomised and two questions were presented to the subjects: a) *"how pleasant do you find this voice based on this sample?"* b) *"would it be easy to listen to this voice for long stretches of time?"*. The results of these tests are shown in figure 3.

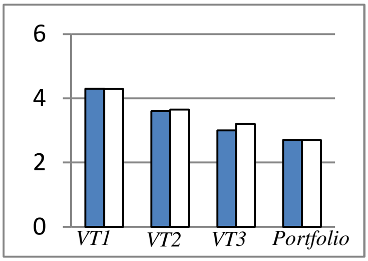

**Fig. 3.** Plot showing MOS evaluation comparing different voice talents and a reference portfolio voice ("Solid fill" bar denotes pleasantness, "no fill" bar denotes the ease of listening). The plot shows a clear preference for VT1 compared to the other voice talents. The speaker and speaking style of the portfolio voice again being least preferred.

A further evaluation using 42 native US subjects was conducted to elicit the prominence of important characteristics in each of the voice talent recordings. Each voice talent recorded 5 passages of 15-20sec. The samples were randomized. In order to streamline the evaluation, the following options were provided to listeners to describe the characteristics of the voice talents: *Spontaneous, Friendly, Lively, Reserved, Competent, Professional, Relaxed, Extravert, Attractive, Successful, Natural, Modern, Sexy, Boring, Slow and Fast.* In addition, the listeners could provide free form description of the audio samples. Figure 4 shows the results for two voice talents, VT1 and the portfolio voice. The results show that the primary characteristics of VT1 are *Friendly, Lively, Professional, Attractive and Natural.* While the primary characteristics for the portfolio voice are *Professional*, *Boring*, *Slow*, *Reserved* and *Competent.* These results nicely summarize the expected response to the voice talent recordings. The portfolio voice has been recorded to meet the demands of traditional customers, looking for a clear professional voice, while also meeting the demands of the technology which require the voice talent to speak in a neutral style. In contrast, the voice talent recordings were designed to meet the needs of conversational agents and come across as friendly and lively. However, in these recordings fewer constraints were placed on the speaker. They were asked to produce a "natural" read reflecting the content of the text. This also comes through in the results.

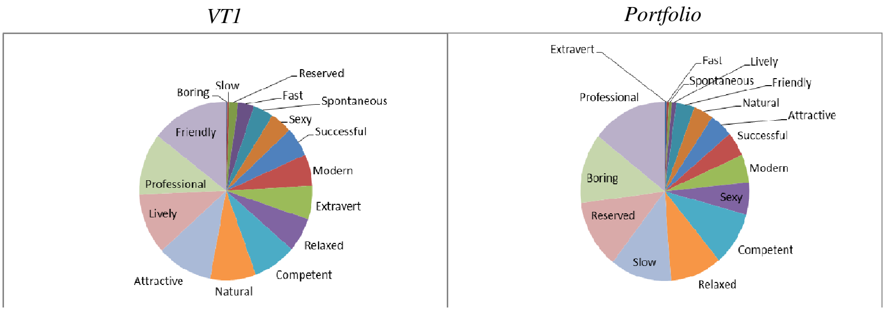

**Fig. 4.** Two pie charts of results for the "characteristics" experiments. The effects of the different speakers and speaking styles are clearly evident.

One final evaluation using 10 native US subjects was conducted. In this experiment two sets of recordings spoken by the voice talent considered to be the best candidate were evaluated. One set was recorded in a neutral style similar to that adopted for portfolio voices and one set recorded using a natural reading style appropriate to the material. Subjects were asked to rate the recordings on listening effort, pleasantness and naturalness.

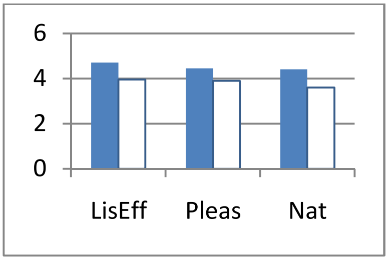

**Fig. 5.** Plot showing the effects of reading style on subject preference. The "solid fill" bars represents natural read, "no fill" bars neutral read.

Figure 5 clearly shows that speaking style affects all three metrics and confirms that the voice talent selection alone is not enough to produce the desired results. The implications of these findings will be discussed in more detail in Sections 4 and 5.

## 3 The Scripts

As shown in Fig. 1, concatenative synthesis uses a database of indexed coded audio samples. The encoding may be simple or complex but in either case, there is a fundamental limitation on the size of the unit inventory, the design of which can have a profound effect on performance and quality. In the absence of a particular customer or target application, scripts tend to focus on two properties in their design. These are basic phonetic coverage and common domains. Basic phone coverage is a fundamental requirement when creating a concatenative TTS system. However, the definition of complete coverage is not as clear as it may first appear to be. Early concatenative systems were designed on the principle of diphone units. The theoretical justification for which was that coarticulation effects between phonemes were captured in the diphone unit, leaving segmental joins to be made in the relatively stable mid-phone point, leading to improved segmental clarity and smoothness. Prosody in such systems was predicted through a prosodic model and applied using signal processing such as PSOLA [2]. The size and completeness of the phone database was determined by the number of phonemes and by basic features such as lexical stress. However, researchers recognised that overall quality could be improved if larger units were stored, and specifically larger units which covered common domains e.g. dates, times. It was also recognised that prosodic models were limiting the naturalness of synthesis systems. Often high quality could be achieved through selecting a mixture of units where prosody was included as a selection feature and not predicted and post applied. These trends lead to what are termed "pure selection" synthesis systems [4,5,6]. Such developments resulted in an explosion in the number of features used to select units and consequently significant growth in the size of the unit database. It also had the effect of breaking the simple definition of database completeness. The growth of features means that even the largest practical database suffers from unit sparsity.

Script design as well as the voice talent and the recording style influence overall acceptability. The more a script can be targeted to the application, the higher the chance of units matching the input can be found in the database and the higher the chance of longer unit sequence being selected leading to improved synthesis.

In order to design a voice to meet the specific demands of conversational agent applications a new approach to creating scripts was considered. This approach is summarised in Fig 6. In this approach conversational agent applications were considered to consist of a series of overlapping domains. These domains were classified into closed and open depending on factors such as the complexity of the language and the likelihood of seeing a large number of out of vocabulary (OOV) items. For example, telephone numbers would be considered a closed domain, as it consists of a well specified syntax and a defined word set. In contrast news is both structurally complex and likely to contain OOV items.

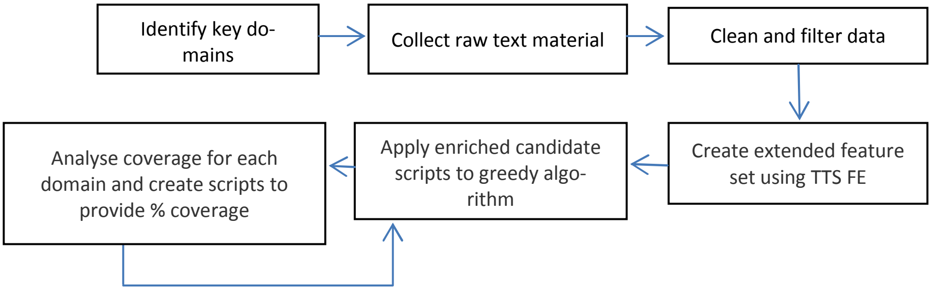
Identify key domains
Collect raw text material
Clean and filter data
Create extended feature set using TTS FE
Apply enriched candidate scripts to greedy algorithm
Analyse coverage for each domain and create scripts to provide % coverage

**Fig. 6.** Steps involved in creating a script optimised for conversational agent applications

Table 1 lists the top domains identified in conversational agent applications. Data from each of these domains was collected and filtered. Finally an iterative script creation process was performed where a “master“ script was created through progressively refining coverage of units in each of the key domains. Using this approach, coverage of key phones and features for each domain could be calculated and combined to provide a master script which was tuned to the overall application.

The script creation process cannot hope to accommodate all the features used in modern unit selection. Some features are assumed to be “implicitly captured“ through the text selection in combination with the recording process. No attempt was made to include fine grained prosodic features as part of script creation. However, in order to capture gross prosodic traits, features which could be robustly extracted from the text were included in the script creation process to supplement the traditional phonemic features. The combination of phonemic and orthographic cues was termed the enriched candidate set. Examples of these orthographic cues are shown in Table 2.

**Table 1.** Top domains identified in conversational agent applications

| Domains | Description |
|---|---|
| Dialogue | General discourse e.g. “hello” |
| Knowledge | “What is a …” |
| Entertainment | “Who is…” |
| Weather | “What is the weather ...” |
| Navigation | “Where is …” |
| Number | “How much…” |
| Calendar | “When is …” |

**Table 2.** Examples of features used to create an enriched set for script creation

| Boundary type | Orthographic cue |
|---|---|
| Document Initial/Final boundary | Degenerate |
| Paragraph Initial/Final boundary | Identified by TTS FE |
| Sentence Initial/Final boundary | . ! ? |
| Within sentence Initial/Final boundary | , ; |
| Parenthetical Initial/Final boundary | () [] {} “ ‘ - _ |

Table 3 shows the percentage of enriched features covered in a specific domain for a pre-defined number of script entries.

**Table 1.** Percentage of enriched features covered in a specific domain for a pre-defined number of script entries

| Weather | Phone cov. | Diphone cov. | Triphone cov. |
|---|---|---|---|
| Maximum phones | 100% | 73.5% | 70.1% |
| Maximum phones and diphones | 100% | 98.4% | 95.8% |
| Maximum phones and triphones | 100% | 97% | 95.6% |
| Maximum diphones | 99.5% | 98.4% | 95.9% |

As described above, domains are defined in terms of whether they are open or closed. This is clearly not a binary classification; rather domains can be seen as having an "openness" property, which can be described in terms of the number of sentences needed to cover a specified number or percentage of enriched features in the scripts. Figure 7 shows the data collected for weather and navigation domains. Navigation has greater openness. This metric is highly dependent on the sampled data. A flowery description of the weather will have very different properties to a terse description. Navigation phrases which do not include place names will have very different properties from a data set which does.

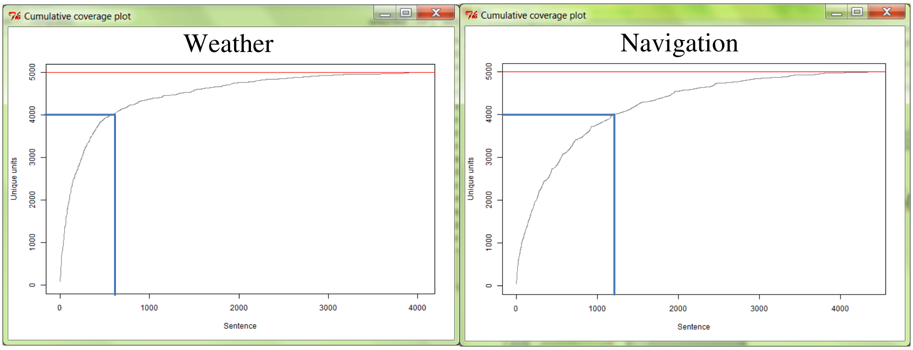

**Fig. 7.** Two diagrams showing how different domains have different "openness" properties

Earlier in this section it was mentioned that for practical computational reasons fine grained prosodic features, while used in the unit selection process during synthesis, are not explicitly considered in script creation. It is assumed that characteristic prosodic patterns will be captured as part of the recording style. This assumption also highlights the issues raised in Section 1, which considered the influence of style on acceptance. The next section considers these prosodic factors in more detail.

## 4 Prosody

Prosody can be viewed as the rhythm, stress and intonation of speech, and is fundamental in communicating a speaker's intentions and emotional state [3]. In TTS systems the prosody prediction component produces symbolic information (e.g. stress patterns, intonation and breath groups) which may or may not be augmented with parametric information (e.g. pitch, amplitude and duration trajectories). Combined, these features are used to define the prosodic realization of the underlying meaning and structure encoded within the text, and are used as feature constraints in unit selection.

There are two fundamental challenges to generating natural synthetic speech; the first challenge is to match the predictive power of the FE with the granularity of labelling of the speech data. The FE must identify and robustly extract features which closely correlate with characteristics observed in spoken language. These same features must be identified and robustly labelled in the unit database. A unit database labelled with too few features matched to a powerful FE will lead to poor unit discrimination during selection, while a weak FE which can only produce a limited set of features will lead to inaccessible units when matched with a richly labelled database. In other words, the expressive power of the FE must match the expressive power of the labelling. The second challenge is that of data sparsity. As already discussed, the unit database is finite, in order to produce high quality synthesis, sufficient examples must exist to adequately represent the expressive power of the features produced by the FE. As prosody is used in selection, the audible effects of sparsity increase as the style of speech becomes more expressive. One way to control these effects is to limit the number and weight of prosodic features. However, such an approach only works well if matched with recordings where prosody is strictly controlled. Weak prosody control during selection when coupled with an expressive database leads to unnatural prosody and segmental "glitching". Another motivation for controlling the style is database size. A neutral prosody will result in a substantially smaller database than one which attempts to capture expressive speech. These two reasons are why the majority of TTS systems strictly control the prosody during recording. Unfortunately, these constraints also limit the acceptability of conversational style synthesis.

Figure 8 shows how expressive speech has greater pitch variability and range compared to a relatively neutral style. This increase must be constrained through selection, while controlling the number of added features, which fragment the search space and exacerbate the problem of unit sparsity. Understanding and controlling sparsity is an active research area [9,10]. As described in Section 3 the traditional features were augmented with additional document and prosody features. An example of the type of symbolic prosodic features considered is shown in Table 4.

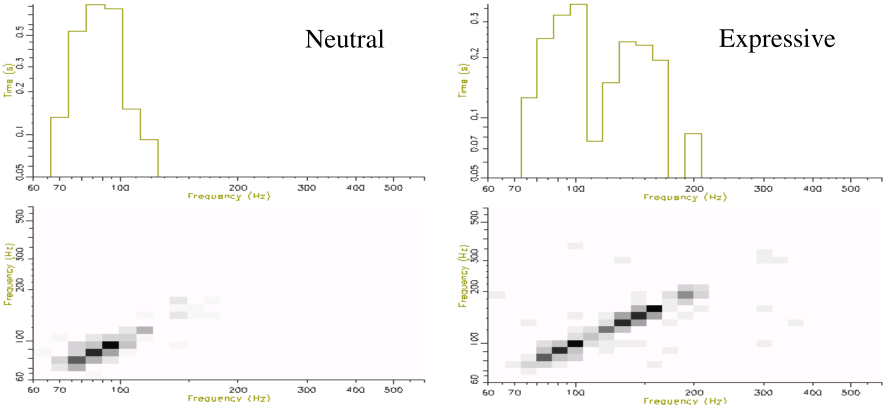

**Fig. 8.** (a) Two plots [8] (neutral style and expressive styles) showing (a) A histogram of all glottal periods (elapsed time, logarithmic scale), (b) a scatter-plot between adjacent glottal periods on a grey-scale.

As previously stated, in order to appreciate the benefit of a richer feature set, good FE prediction and accurate labelling must go hand in hand. A method of automatically labelling the speech data with the type of features shown in Table 4 has been produced, and matched with a "lazy learning" technique for prediction. Accuracy rates of above 75% across all the features investigated have been achieved.

**Table 2.** Description of prosodic features produced by the FE

| Label | Description |
|---|---|
| Word Prominence Level | Reduced: typically function words, no lexical stress, no pitch movement.<br>Stressed: stressed syllable in (content) word.<br>Accented: stressed syllable in (content) word has salient pitch movement.<br>Emphasized: stronger than accented. |
| Prosodic Phrase Boundary | Word<br>Weak: intonation phrase.<br>Strong: Major phrase. |
| Sentence | Phrase type (Prototypical Phrase Intonation Contour). |

This section and the previous sections have described how a speech database tailored for conversational agent applications has been recorded, designed, and labeled. The next section describes the synthesis method used to produce smooth and compelling synthesis using this data.

## 5 Synthesis

Expressive speech requires a synthesis process complex enough to control the greater prosodic variability found. Nuance has been working for many years on a method of selection and synthesis capable of supporting expressive speech. The method, called Multi-form synthesis (MFS) is a statistically motivated hybrid approach which combines the segmental quality benefits of concatenative systems with the flexibility and trainability of model based approaches. A detailed description of this approach can be found in [11]. Hybrid methods have been shown to be robust to sparsity which, as discussed above, is one of the side effects of expressive speech. However in order to produce compelling results, MFS must be combined with the rich prosody prediction discussed in Section 4. Without such prosodic control, synthetic speech may sound smooth but with unnatural prosody.

Figure 9 diagrammatically shows the key processes of MFS synthesis. In this diagram, input text is analysed to create feature vectors. These may be complex features as described in Section 4. A search is then performed matching the phonetic and prosodic context vectors to the HMMs model in inventory *M*, from which a sequence of model segments *Om* is obtained. These model segments are used to direct a search of template candidates $O_T$ in the template inventory T. *Om* is also used to generate (Pgen PI) a first set of parameter trajectories *p*.

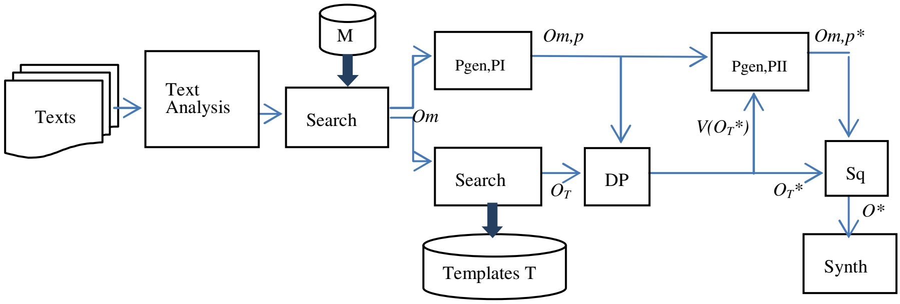

**Fig. 9.** Multi-Form Synthesis (MFS)

The model segments, parameter trajectories and template candidates are input into a dynamic programming algorithm (DP). As a result, the best template segment sequence $O_T$* is obtained. The variance of parameter trajectories of the best template sequence is fed into a second parameter generation algorithm (Pgen PII) which regenerates the parameters *p**. The result of this process is a sequence of speech parameter trajectories *Om,p** with variance reassembling the variance of the best template segments. This is done to combine seamlessly template segments with model segments. Finally, the best models and template segments are sequenced. This sequence *O** of "multiform" segments is sent to the synthesizer-concatenator. The parameter trajectories are converted to synthetic speech and concatenated with the template segments, which yields the speech output waveform.

## 6 One Last Trick

So far this paper has concentrated on how to create material for expressive voices and how to use this material within a TTS system. As previously mentioned, there are limits to the degree of expressivity which can be accommodated within a TTS system, even one designed to support expressive speech. In addition, para-linguistic sounds such as laughing, crying, exclamations etc. do not fit easily into traditional linguistic analyses. Fortunately there is a simple pragmatic approach which can be used to support highly expressive and para-linguistic elements. In this approach, shown in figure 10, key idiomatic phrases ("gilded phrases") are recorded and sit alongside traditional unit selection synthesis. During synthesis, orthographic pattern matching is used to identify fragments in the text. When such fragments are identified, a gilded phrase (pre-recorded phrase fragment) is selected instead of a full FE analysis and BE synthesis. Such an approach can be highly effective for domains such as dialogue prompts which consist of frequently re-occurring highly expressive phrase patterns.

Gilded phrases can be identified as separate elements during script design, or as shown in Fig. 10, they can be automatically constructed from an analysis of the standard script elements and used to supplement or augment the main unit inventory.

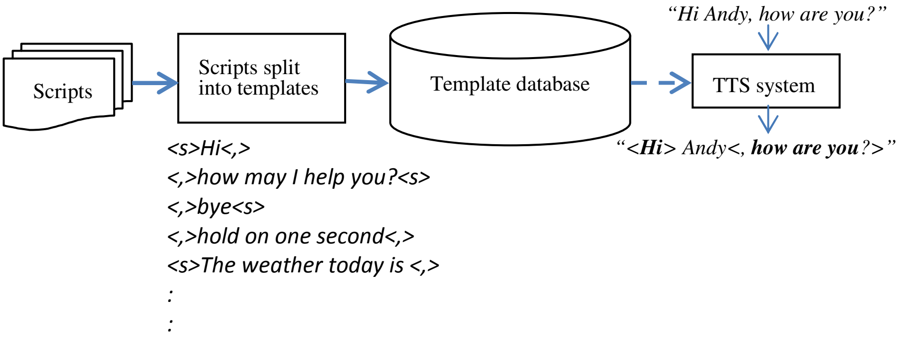
"Hi Andy, how are you?"
Scripts
Scripts split into templates
Template database
TTS system
<s>Hi<,>
<,>how may I help you?<s>
<,>bye<s>
<,>hold on one second<,>
<s>The weather today is <,>
:
:
"<Hi> Andy<, how are you?>"

**Fig. 10.** Diagram showing the construction and use of "gilded phrases"

## 7 Results and Conclusions

This paper has focused on the creation of a specific type of speech synthesis, expressive conversational speech. The early part of the paper demonstrated the importance of matching the recoding talent and style to the target domain. The later sections described why expressive speech places additional demands on traditional concatenative systems, and briefly described how Nuance addresses these challenges. Figure 11 shows the results of a MOS evaluation which compared our latest expressive system with a reference conversational agent application. It can be seen that the new system outperforms the reference both for a closed domain and an open domain.

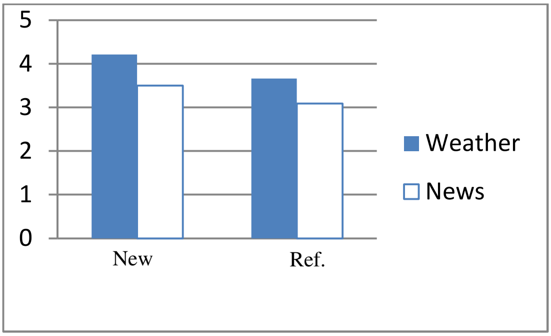

**Fig. 11.** MOS evaluation against a reference system

**Acknowledgments.** The work described in this paper is not solely the work of any one individual but represents the combined efforts of the Nuance TTS R&D team. I would like to take this opportunity to thank the team for all their hard work and dedication.

## References

1. Klatt, D.: Review of text-to-speech conversion for English. J. Acous. Soc. Amer. 82, 737–793 (1987)
2. Taylor, P.: Text-To-Speech Synthesis. Cambridge University Press (2009)
3. Ladd, D.R.: Intonational Phonology. Cambridge University Press (1996)
4. Breen, A.P.: The BT Laureate Text-To-Speech System. In: ESCA/IEEE Workshop on Speech Synthesis, pp. 195–198 (1994)
5. Hunt, A., Black, A.: Unit selection in a Concatenative Speech Synthesis System using a Large Speech Database. In: ICASSP, pp. 373–376 (1996)
6. Donovan, R.: Trainable Speech Synthesis, PhD Thesis, University of Cambridge (1996)
7. Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T.: Simultaneous Modelling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. In: Eurospeech 1999, pp. 2374–2350 (1999)
8. SFS "Speech Filing System", `http://www.phon.ucl.ac.uk/resource/sfs/`
9. Chen, L., Gales, M.J.F., Wan, V., Latorre, J., Akamine, M.: Exploring Rich Expressive Information from Audiobook Data Using Cluster Adaptive Training. In: Interspeech 2012 (2012)
10. Zen, H., Senoir, A., Schuster, M.: Statistical Parametric Speech Synthesis using Deep Neural Networks. In: ICASSP, pp. 7962–7966 (2013)
11. Pollet, V., Breen, A.P.: Synthesis by Generation and Concatenation of Multi-form Segments. In: ICSLP 2008 (2008)