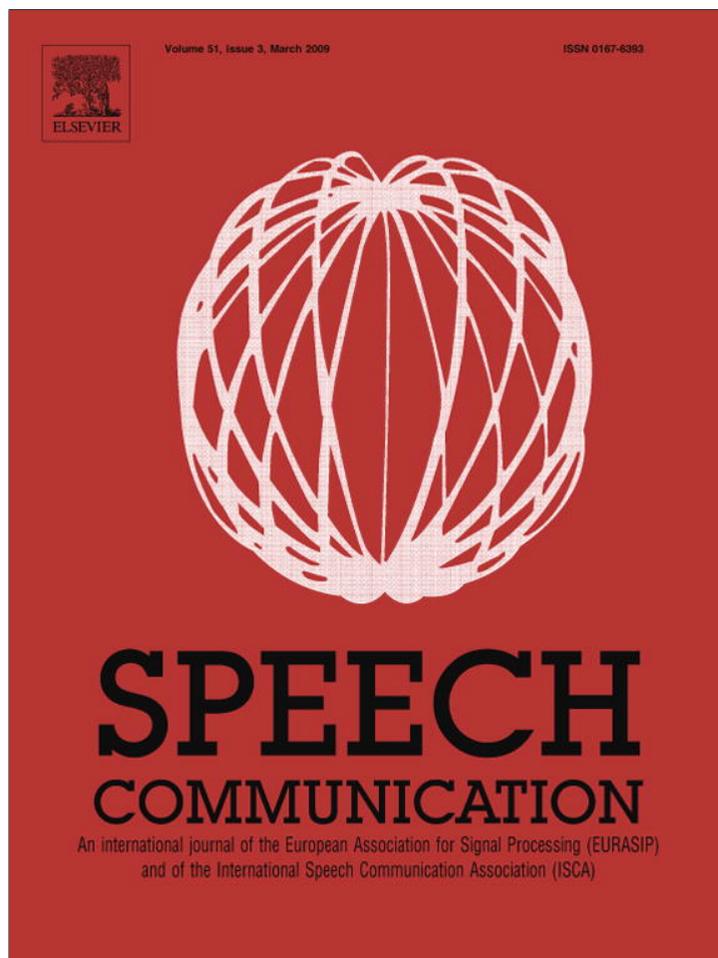


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Data-driven emotion conversion in spoken English

Zeynep Inanoglu, Steve Young*

University of Cambridge, Department of Engineering, Trumpington Street, Cambridge CB2 1PZ, UK

Received 29 May 2008; received in revised form 9 September 2008; accepted 9 September 2008

Abstract

This paper describes an emotion conversion system that combines independent parameter transformation techniques to endow a neutral utterance with a desired target emotion. A set of prosody conversion methods have been developed which utilise a small amount of expressive training data (~15 min) and which have been evaluated for three target emotions: anger, surprise and sadness. The system performs F0 conversion at the syllable level while duration conversion takes place at the phone level using a set of linguistic regression trees. Two alternative methods are presented as a means to predict F0 contours for unseen utterances. Firstly, an HMM-based approach uses syllables as linguistic building blocks to model and generate F0 contours. Secondly, an F0 segment selection approach expresses F0 conversion as a search problem, where syllable-based F0 contour segments from a target speech corpus are spliced together under contextual constraints. To complement the prosody modules, a GMM-based spectral conversion function is used to transform the voice quality. Each independent module and the combined emotion conversion framework were evaluated through a perceptual study. Preference tests demonstrated that each module contributes a measurable improvement in the perception of the target emotion. Furthermore, an emotion classification test showed that converted utterances with either F0 generation technique were able to convey the desired emotion above chance level. However, F0 segment selection outperforms the HMM-based F0 generation method both in terms of emotion recognition rates as well as intonation quality scores, particularly in the case of anger and surprise. Using segment selection, the emotion recognition rates for the converted neutral utterances were comparable to the same utterances spoken directly in the target emotion. © 2008 Elsevier B.V. All rights reserved.

Keywords: Emotion conversion; Expressive speech synthesis; Prosody modeling

1. Introduction

The ability to output expressive speech via a Text-to-Speech Synthesiser (TTS) will make possible a new generation of conversational human–computer systems which can use affect to increase naturalness and improve the user experience. Typical examples include call centre automation, computer games, and personal assistants.

To avoid building a separate voice for each required emotion, a transformation can be applied to modify the acoustic parameters of neutral speech such that the modified utterance conveys the desired target emotion. However, learning the complex rules that govern the

expression of any target speaking style is a significant challenge and although various rule-based transformation attempts exist in the literature (see [Schroder, 1999](#) for a review), designing good rules for each expressive style requires tedious manual analysis and even then, only a very limited set of acoustic-prosodic divergences can be captured.

In this paper we explore a set of data-driven emotion conversion modules which require only a small amount of speech data to learn context-dependent emotion transformation rules automatically. The data-driven conversion of acoustic parameters in speech has been widely-studied in the field of voice conversion. Whilst conceptually emotion conversion can be thought of as a form of voice conversion, in practice, voice conversion techniques have focused on the transformation of the vocal tract spectra, and relatively little attention has been paid to adapting long-term

* Corresponding author. Tel.: +44 223 332654; fax: +44 223 332662.
 E-mail addresses: zeynep@gatesscholar.org (Z. Inanoglu), sjy@eng.cam.ac.uk (S. Young).

F0 and duration patterns (see Stylianou et al., 1998; Ye, 2005; Kain and Macon, 1998). For example, a popular F0 transformation technique employed in conventional voice conversion is Gaussian normalization, which scales every pitch point in the source speaker's F0 contour to match the mean, μ_t and standard deviation, σ_t of the target:

$$F(s) = \frac{\sigma_t}{\sigma_s} s + \mu_t - \frac{\sigma_t \mu_s}{\sigma_s}, \quad (1)$$

where μ_s and σ_s are the mean and standard deviation of the source.

More complex F0 conversion functions have been proposed for voice conversion: Inanoglu (2003) investigated GMM based F0 transformation, Gillett and King (2003) performed piecewise linear transformation based on salient points in the contour and Helander and Nurminen (2007) used a codebook-based approach to predict entire F0 segments using linguistic information. However, these methods have mainly been designed and evaluated within the context of speaker conversion where the focus is on transforming the prosodic characteristics of one speaker to sound like another. In this scenario, the speech is typically neutral and exhibits minimal prosodic variability.

Due to the dominant role of F0 and duration patterns in distinguishing emotional expressions (Yildirim et al., 2004; Barra et al., 2007; Vroomen et al., 1993), the focus of this paper will be on the transformation and evaluation of prosody within an emotion conversion framework. Scherer and Banziger (2004) demonstrates that emotions can have a significant effect on global F0 statistics such as mean and range as well as F0 contour shapes. In an effort to capture both kinds of effects within a single framework, we adopt a linguistically motivated approach to emotion conversion, by making explicit use of text-based linguistic details as predictors in our transformation methods. A recent study by Bulut et al. (2007) which attempts to analyze the interaction between part of speech tags, sentence position and emotional F0 contours support this modeling approach.

Various methods of emotion conversion have been reported in the literature. In (Kawanami et al., 1999), GMM-based spectral conversion techniques were applied to emotion conversion but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. Wu et al. (2006) proposed a unified conversion system using duration embedded Bi-HMMs to convert neutral spectra and decision trees to transform syllable F0 segments. In (Tao et al., 2006), the use of GMM and CART-based F0 conversion methods were explored for mapping neutral prosody to emotional prosody in Mandarin speech. Data-driven emotion conversion methods specifically for use in an HMM-based speech synthesis framework have also been implemented by Tsuzuki et al. (2004) and Yamagishi et al. (2003).

In this paper we describe an emotion conversion system for English which can be used to add an additional layer of expressiveness to an existing speech synthesizer without sacrificing quality. In such a scenario, the waveform output

of any synthesis system, regardless of the underlying synthesis technique, can be the input for emotion conversion. An overview of the conversion system is illustrated in Fig. 1 and its modules can be summarized as follows:

- **Spectral Conversion:** As the first step, spectral conversion is performed at the level of pitch-synchronous speech frames. LPC-based analysis steps are used to extract a train of vocal tract features for conversion. A GMM-based linear transformation method is applied to this feature sequence to change the neutral voice quality to that of a target emotion. This method is based on the earlier technique of (Ye, 2005), which was used for speaker identity conversion. Finally, Overlap and Add (OLA) synthesis is then used to resynthesise a waveform with the desired emotional voice quality combined with the original neutral prosody.
- **Duration Conversion:** The input to duration conversion consists of neutral phone durations for a given utterance as well as the linguistic context of each phone. A set of regression trees specific to each broad phone class are then used to transform neutral phone durations for a given target emotion, resulting in a duration tier. The duration tier, which is simply a sequence of duration scaling factors for an input utterance, is then used by the Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) implementation provided by the Praat software¹ (Boersma and Weenink, 2005) to scale phone durations in the waveform.
- **F0 Conversion:** Two alternative methods for F0 generation are compared within the full-conversion system. Both methods require syllable boundaries as input as well as the linguistic context of the syllable. If duration conversion is performed in the previous step, the new syllable boundaries are computed and input into the selected F0 generation module. The transformed F0 contour for an utterance is transplanted onto the waveform using TD-PSOLA.
 - **HMM-based F0 Generation:** Context-sensitive syllable HMMs are used to model and generate expressive F0 contours. Syllable representation is based on a small pool of linguistic features which are acquired from the text-based representation of the utterance.
 - **F0 Segment Selection:** As in unit-based speech synthesis, this method is based on a concatenative framework where syllable F0 segments from an expressive corpus are combined to form new utterance contours in a way that minimizes a cost function. In this context, a syllable F0 segment refers to that part of an utterance F0 contour which falls within the boundaries of a given syllable. The unvoiced regions within the syllable do not contribute to the F0 segment definition. Given a sequence of such syllable F0 segments for a neutral utterance

¹ <http://www.fon.hum.uva.nl/praat/>.

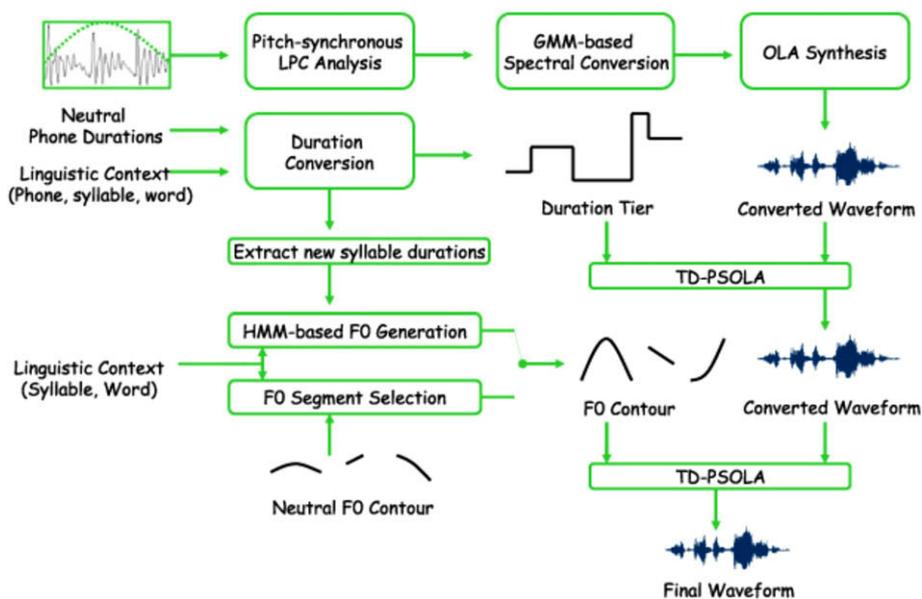


Fig. 1. Overview of emotion conversion system.

and their corresponding syllable contexts, the goal of segment selection is to replace these segments with emotional counterparts by performing a dynamic search through a trellis of available candidates in the target corpus.

In this paper we first describe a set of preference tests to individually evaluate each module in terms of its contribution to emotion perception. Some care is needed in interpreting these results since the modules may have complex interactions when they are combined. Barra et al. (2007), for instance, found that angry prosody resulted in sad sounding speech when combined with neutral spectra. When combined with angry spectra, however, angry prosody was able to enhance the perception of anger compared to case where only the angry spectra were evaluated. This motivates our decision to always include spectral conversion when contrasting the performance of the different F0 conversion modules. Finally, a further emotion classification test was conducted on the complete system to evaluate the overall success in conveying the desired emotions.

The rest of the paper is organized as follows: in Section 2, details of emotional speech data collection are reported. Section 3 describes the HMM-based F0 generation method and Section 4 presents F0 segment selection as an alternative. In Section 5, duration conversion is explained, while Section 6 provides an overview of the spectral conversion module. Finally, in Section 7, the results of the perceptual listening tests are reported.

2. Emotional speech data

The emotional speech data used in this work was recorded as part of a wider data collection effort organized

by the Speech Technology Group, Toshiba Research Europe. A professional female voice talent recorded parallel speech data in three expressive styles (angry, surprised, sad) as well as a neutral style. While only three emotions were used as case studies, the methods proposed in this paper could be applied to any other target expressive style which shows consistent acoustic behavior. Our choice of three target emotions is motivated by the diversity of their acoustic profiles: anger is known to have a dominant spectral component, while surprise is mainly conveyed through prosody and sadness is thought to be conveyed by both (see Barra et al., 2007).

In expressing the emotions, the voice talent was asked to assume a natural, conversational style rather than a dramatic intensity. A total of 300 utterances were recorded for each emotion using prompt scripts extracted from the standard unit-selection corpus used to train a commercial TTS system. The sentences in this subset were chosen to preserve phonetic coverage. Of the 300 utterances, 272 were used for training and 28 were reserved for testing. This training set size is comparable to that used in other voice conversion studies. For example, it is similar to the emotion conversion experiments in (Wu et al., 2006) and smaller than the emotional prosody conversion experiments described in (Tao et al., 2006). The numbers of words, phones and syllables in the training and test sets are given in Table 1. The mean word count per sentence is 7.9. The total duration of speech data used for training was around 15 min for each emotion.

Table 1
Number of linguistic constituents in training and test sets

| | Utterances | Words | Syllables | Phones |
|-----------------|------------|-------|-----------|--------|
| Training corpus | 272 | 2170 | 3590 | 10,754 |
| Test corpus | 28 | 215 | 367 | 1115 |

For each data file in the corpus and its word level transcription, a set of corresponding annotation files were automatically generated. Phone and state boundaries were extracted using HTK-based forced alignment (Young et al., 2006). The Cambridge English Pronunciation dictionary was used to identify syllable constituents for each word, as well as lexical stress positions (Jones, 1996). A proprietary part-of-speech tagger was used to mark each word with one of 16 part-of-speech tags. Based on these extracted linguistic features and the boundary information, a hierarchical computational map of each utterance was built in preparation for processing by the conversion modules. Pitch contours and pitch marks were also extracted directly from the waveform using Praat software¹ (Boersma and Weenink, 2005) and manually corrected for mistakes.

3. F0 generation from syllable HMMs

HMMs have been used for the recognition of F0 related features such as accent and tone for some time (see Fallside and Ljolje, 1987; Jensen et al., 1994). However, the use of HMMs as generators is more recent, and is mostly due to the development of HMM-based synthesis technology. The most popular HMM-based speech synthesis system, HTS, proposed by Tokuda et al. (2002), allows simultaneous modeling and generation of F0 and speech spectra for full-spectrum speech synthesis as long as a significant amount of data is available to train the phone models. The appropriateness of phone models for modeling F0 contours in isolation, however, is arguable, since most prosodic labeling systems such as Tones and Breaks Indices (TOBI) consider the syllable to be the fundamental building block of intonation (Silverman et al., 1992). The method described here adheres to this convention by modelling F0 at the syllable level based on features derived from syllable and word level transcriptions. Of specific interest is the interaction between syllable and word level linguistic identifiers and emotional F0 contour shapes, an area “largely unexplored” according to a study published by Banziger and Scherer (2005).

3.1. Model framework

The starting point of our models is the association of syllables with their corresponding F0 movements. Unlike phonetic symbols, syllables do not have a widely-accepted symbol or label which provides a link to F0 movements. Pitch accent classification schemes such as TOBI have been used to model syllable-based F0 in neutral speech (see Inanoglu and Young, 2005; Ross and Ostendorf, 1994). However, TOBI-derived units are far from ideal, since they require manual labeling of training data by expert humans and even then they manifest high inter-labeler disagreement.

In this paper, we explore a set of text-based syllable and word-level linguistic features that are common to all emo-

tional renderings of a given utterance. The choice of features used here resulted from a literature review and informal listening tests. Priority was given to features that are readily available in a TTS context.² The features we used are position in sentence (spos), lexical stress (lex), position in word (wpos), part of speech of current word (pofs), part of speech of previous word (ppofs), onset type (onset) and coda type (coda). Table 2 illustrates an example sentence and the syllable labels for the last word “sloppy”. Word position is identified explicitly for the first three words and the last three words in the sentence. All words in between these sentence-initial and sentence-final word groups are identified with a single tag (spos = 4), resulting in a total of seven position values. For example, both syllables of “sloppy” take the position value 7 (spos = 7) indicating that they belong to the last word of the utterance, while all syllables in the words “a”, “forage” and “cap” would take the position value 4 (spos = 4). Lexical stress is either set as 1 for stressed or 0 for unstressed. Syllable position in the word can have four values: beginning of word (wpos = 1), middle of word (wpos = 2), end of word (wpos = 3) or a value indicating a single-syllable word (wpos = 0). Thirteen part-of-speech tags were used based on a proprietary part-of-speech tagger. In the example of “sloppy”, part of speech is set to 6 to identify an adjective and previous part-of-speech is set to 2 to identify a verb. Finally, two additional intra-syllable features, onset and coda, were used in order to provide clues about the basic phonetic structure of the syllable. We define onset as the consonant that precedes the syllable nucleus. This phone can be non-existent (onset = 0), unvoiced (onset = 1) or sonorant (onset = 2). The same categorical distinctions are also made for the syllable coda, the consonant sounds that follow the nucleus.

The training data contains 2086 unique combinations of features. Table 3 summarizes the percentage of feature combinations in the test data which have not been observed in the training data. Although 42.3% of the combinations in the test set are unseen, matching only 6 of the 7 features reduces the unseen combinations to 2.8%, indicating that for almost all the test data, a very similar if not exact context has been observed in the training data. This motivates the use of decision-tree based parameter tying where unseen contexts can be modeled using similar alternatives.

3.2. Model training

In order to model the mix of voiced and unvoiced speech in the F0 stream, Multi-Space Probability Distribution HMMs (MSD-HMMs) were adopted as described in (Tokuda et al., 2002). The voiced region within each syllable was aligned with the context-dependent syllable labels determined by the corresponding linguistic features. The

² A detailed search for an optimal feature set which maximizes emotion perception for a given emotion is an interesting area but beyond the scope of this paper.

Table 2

Syllable labels for a two-syllable word “sloppy” which appears in the sentence “A soldier in a forage cap always looks sloppy.” Contextual key-value pairs are separated by a colon

| A soldier in a forage cap always looks SLOPPY | | |
|---|--------|--|
| Syllable | Phones | Context Labels |
| 1 | s:l:aa | spos = 7:lex = 1:wpos = 1:pofs = 6: ppofs = 2:onset = 2:coda = 0 |
| 2 | p:ii | spos = 7:lex = 0:wpos = 3:pofs = 6: ppofs = 2:onset = 1:coda = 0 |

Table 3

Percent of unseen contexts in the test data

| | Number of matching features | | |
|---------------------|-----------------------------|------------|------------|
| | 7 features | 6 features | 5 features |
| Unseen contexts (%) | 42.3 | 2.8 | 0 |

unvoiced regions in the training utterances were modeled using a separate *uv* model. Fig. 2 illustrates an example of label alignments for a short speech segment. It is important to note that the actual syllable boundaries shown in Fig. 2a and b are modified to include only the syllable voiced segments in Fig. 2c, where the unvoiced *uv* labels have been inserted.

The F0 model training follows a conventional recipe analogous to that provided by the HTS system.³ Three state left-to-right HMMs are used with three mixture components where two of the mixtures represent the continuous voiced space and the third represents the discrete “unvoiced” space. The input feature stream consists of F0 values as well as their first and second order differentials. A separate set of syllable models were built for each of the three emotions: surprised, sad and angry.

In speech recognition and HMM-based speech synthesis, context-independent monophones are traditionally used for initialization and then, once trained, they are cloned to form the required context-dependent model set. However, in the case of syllable F0 models, a core set of labels analogous to phones does not exist. Hence, in this case, each model is initialised using a subset of the features chosen to ensure a relatively balanced coverage per model. This subset comprised word position in sentence (spos), syllable position in word (wpos) and lexical stress (lex). This feature subset resulted in 56 base models plus a *uv* model for unvoiced regions. The average number of training samples per syllable model was 64. Full-context models were then built by replicating the base models and using further iterations of embedded training. Due to sparsity of data and the fact that a wide range of feature combinations are unseen, decision-tree based clustering was performed based on a log-likelihood criterion. Trees were built for each position in the sentence, and the initial, middle and final states were clustered separately. The initial set of 6258 states (2086 models \times 3 states) were thereby

reduced to 801, 779 and 529 clusters for surprise, anger and sadness, respectively.

3.3. Generation from syllable HMMs

To generate an F0 contour, the required syllable label sequence for a test utterance is derived from its orthographic transcription and syllable boundaries are either copied from the neutral utterance or derived from the neutral utterance using the duration conversion module described below. The parameter generation framework of the HTS system was used in the mode where the state and mixture sequences are known (see Tokuda et al., 2000), the former being determined from the syllable boundaries and the duration models as described in (Yoshimura et al., 1998). The mixture component with the highest weight is used for generation. Once generated, the F0 stream can then be transplanted onto the neutral utterance for perceptual evaluation.

The generative capacity of the trained F0 models is illustrated in Fig. 3, which displays F0 contours generated by the different emotion models for the same syllable label sequence. The full-context label sequence was extracted from the test sentence “Earwax affects koala bears terribly” which consists of 12 syllable labels. The contours clearly display the inherent characteristics of the three emotions: sadness follows a slightly monotone shape with a tight variance; surprise and anger share some characteristics in the beginning of the sentence while at the final voiced segment, a sharp fall is generated for anger, and rising tone for surprise, signaling a question-like intonation which is a common indicator of disbelief.

Finally, over-smoothing of the feature space is a known shortcoming of HMM-based parameter generation. A method has been proposed by Toda and Tokuda (2005) to generate parameters based not only on the log likelihood of the observation sequence given the models but also on the likelihood of the utterance variance which is referred to as global variance (GV). A single mixture Gaussian distribution is used to model the mean and variance of utterance variances. This model is trained separately for each emotion and then integrated into the parameter generation framework. When applied to our syllable F0 models, GV made a perceptual improvement in the case of surprise by increasing the global variance of the generated F0 contour while keeping the contour shape intact. An example of this effect is illustrated in Fig. 4. In the case of anger and sadness, however, the addition of global variance to the

³ HTS Version 2.1 alpha was used.

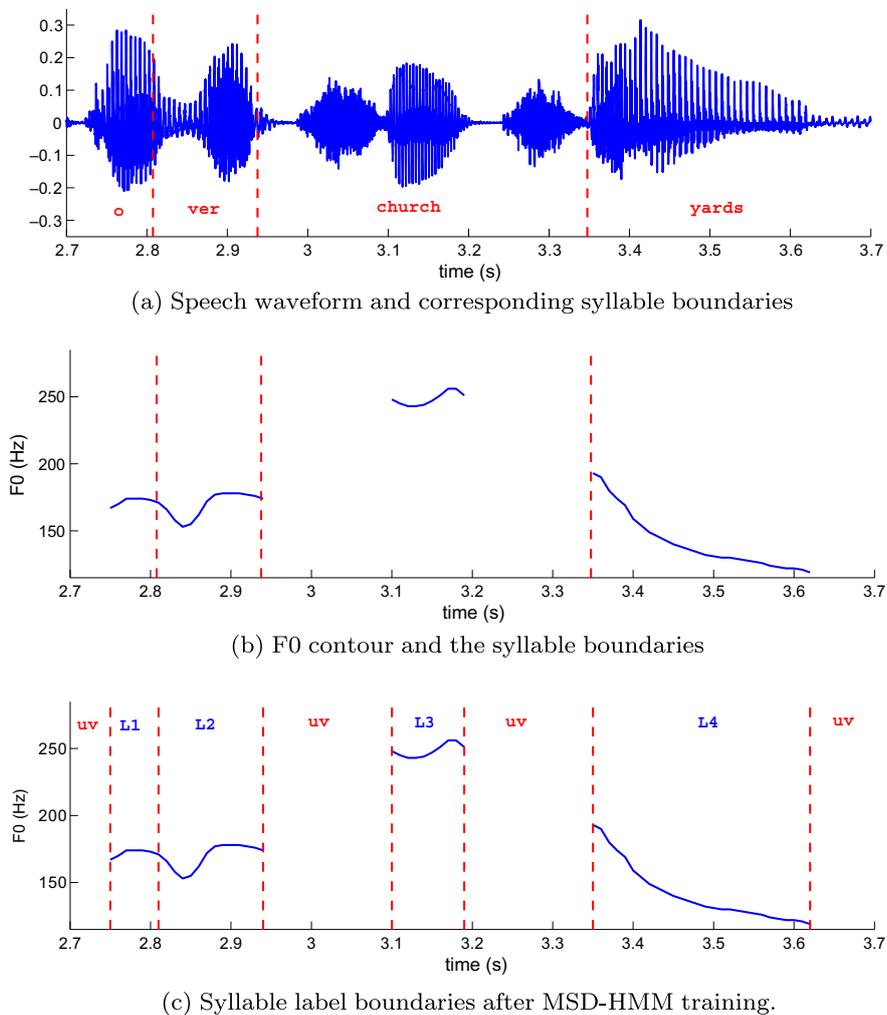


Fig. 2. Example syllable alignment for the phrase “over churchyards”. Labels L1, L2, L3, and L4 represent linguistic feature combinations. (a) Speech waveform and corresponding syllable boundaries; (b) F0 contour and the syllable boundaries and (c) syllable label boundaries after MSD-HMM training.

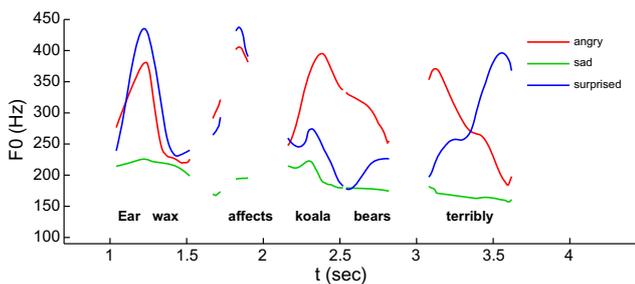


Fig. 3. HMM-generated contours for three emotions using the same utterance “Earwax affects koala bears terribly”.

parameter generation framework made no perceptual difference. This may be due to the fact that HMM parameters for surprise result in over-smoothing of F0 contours, while for anger and sadness the models themselves are sufficient in producing the appropriate variance for the utterance. Hence, in the perceptual evaluations of HMM-based F0 contours described below, GV-based parameter generation is only used in the case of surprise.

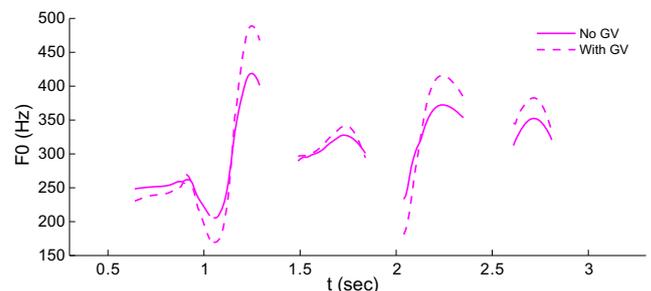


Fig. 4. Surprised F0 contour generated by the syllable models for the utterance “Dry black-thorn is grim stuff” with and without Global Variance (GV) based parameter generation.

4. F0 segment selection

F0 segment selection makes use of a concatenative framework similar to unit selection. A sequence of syllable F0 segments are selected directly from a small expressive corpus, using target and concatenation costs. A similar

idea has been explored by Tian et al. (2007) to predict F0 contours in a non-expressive TTS framework from a large corpus of Mandarin speech. The goal of the method described here, however, is to generate expressive prosody from limited data in a conversion framework. Parallel neutral and emotional utterances are chopped into their corresponding syllables and the voiced F0 segments within these parallel syllables are extracted and stored as part of the unit definition as well as their common linguistic context features. We define a syllable target cost T and an inter-syllable concatenation cost J such that the total cost over S syllables for a given unit sequence U and input specification sequence I is defined as

$$C(U, I) = \sum_{s=1}^S T(u_s, i_s) + \sum_{s=2}^S J(u_{s-1}, u_s). \quad (2)$$

The target cost T is a weighted Manhattan distance consisting of P subcosts

$$T(u_j, i_s) = \sum_{p=1}^P w_p T_p(u_j[p], i_s[p]). \quad (3)$$

Eight target subcosts ($P = 8$) are used. The first seven are binary subcosts indicating whether the individual context features (e.g. lexical stress) in the specification match the corresponding syllable context in the unit. A matching feature results in zero cost whereas a mismatch results in a unit cost of 1. The context features in this section are the same as the ones we have introduced in Section 3 in order to provide a fair comparison of the two methods given the feature set. The final subcost, T_{f0} , is the Root Mean Squared (RMS) distance between the contour $F0^i$ of the input syllable being converted and the neutral contour, $F0^n$, which is stored as part of the unit definition

$$T_{f0} = \sqrt{\frac{1}{L} \sum_{l=1}^L (F0^i(l) - F0^n(l))^2}, \quad (4)$$

where L is the length after interpolating the two segments to have equal duration. It is important to note that alternative formulations of this subcost (e.g. perceptually correlated distance measures) are possible within this framework. In this paper we evaluate the use of RMS error to provide a baseline for future experiments.

The weights for each subcost serve two functions: firstly they normalize the different ranges of categorical and continuous subcosts and secondly they rank features according to their importance for each target emotion.

The concatenation cost, J , is nonzero if and only if consecutive syllables in the input specification are “attached”, i.e. within the same continuous voiced region. If the voiced syllable segment for the input specification i_{s-1} ends at time t_1 and the input specification i_s begins at time t_2 , the concatenation cost for two candidate segments in the target corpus with lengths, L_{s-1} and L_s , is defined as the difference between the last F0 point in segment $F0_{s-1}$ and first F0 point in segment $F0_s$ iff t_1 is equal to t_2 :

$$J(u_{s-1}, u_s) = \begin{cases} w_J (F0_{s-1}[L_{s-1}] - F0_s[1]) & \text{if } t_1 = t_2, \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

The concatenation cost is included to avoid sudden segment discontinuities within voiced regions. A concatenation weight, w_J is used to prioritize this cost relative to the target subcosts when selecting segments.

Once all the costs are defined, segment selection becomes a problem of finding the path, \hat{u} , with the smallest cost through a trellis of possible F0 segments given an utterance specification. Viterbi search is used to find the minimum-cost path, by tracing back locally optimal candidates. Note that the concatenation cost is zero for all syllable voiced segments that are detached from the preceding voiced segments due an intervening unvoiced region or a pause. Therefore if an input utterance consists of only detached syllables, the concatenation cost plays no role in segment selection and the optimal path will simply be the sequence of units which minimize target costs locally at each syllable time step.

Weights for the subcosts are estimated separately for attached and detached syllables. This distinction is motivated by the fact that all weights for target subcosts are likely to change when a concatenation cost exists (i.e. the syllable is attached to its left context). Therefore, two sets of weights are estimated on held-out utterances using the least squares linear regression method described below.

4.1. Weight estimation for detached syllables

For the detached syllable case, a set of P weights, w_p^T , are estimated for each target subcost. For each held out syllable F0 segment in the target emotion, the N -best and N -worst candidates in the corpus are identified in terms of their RMS distance to the held-out segment. This choice emphasizes the units we most want our cost function to select and the units we most want it to avoid. The cost functions for these syllable segments are then set equal to their RMS distances, which results in a system of linear equations which can be solved to yield the required weights. In this framework, the N -worst candidates are useful for capturing the context mismatches that are frequently associated with high RMS errors and which should therefore be given high weights. The N -best candidates, on the other hand, allow us to identify mismatches that are associated with low RMS values and which are probably not important perceptually and hence deserve a low weight. Combining the equations for each of the M held-out syllables and $2N$ candidates yields the following system of $2NM$ equations which can be solved using least squares:

$$CW = D, \quad (6)$$

where C is the $2NM \times P$ matrix of subcosts which have been computed. W is the $P \times 1$ vector of unknown weights and D is the $2NM \times 1$ vector of RMS distances. In our system N was set to 5 and leave-one out cross-validation was

performed on all training utterances to obtain the final system of equations. Informal listening tests with values of N larger than 5 did not show any perceptual improvement. However, more formal tests with a range of values for N should be carried out to confirm these findings. The weights obtained for detached syllables are listed in Table 4. The different contextual weights indicate which features are most relevant for each target emotion. Lexical stress (lex) and syllable position in word (wpos) result in the highest categorical weights across all emotions, indicating that a mismatch in these categories should be strongly penalized. Position in sentence (spos), on the other hand, seems to be one of the least important categorical features for anger and sadness, whereas for surprise it ranks higher. For anger, part of speech (pofs) and previous part of speech (ppofs) seem to be the most important features after lexical stress and word position. The similarity of the input segment to a neutral segment in the corpus also has a dominant effect on segment selection for this emotion ($w_{F0} = 1.00$). This implies a more linear and regular relationship between neutral and angry segment pairs than is the case with surprise or sadness. Note that the low values for the weights w_{F0} is due to the higher mean of the subcosts T_{f0} compared to the categorical subcosts.

4.2. Weight estimation for attached syllables

As noted above, a different set of target weights, w_p^J , are applied to segments that are attached to their left-context, along with an additional weight for the concatenation cost, w_J . From (2) and (3), the local cost of attaching a unit u_k to a preceding unit u_j during selection is

$$C(u_k, i_s) = \sum_{p=1}^P w_p^J (T_p(i_s[p], u_k[p])) + w_J J(u_j, u_k). \quad (7)$$

For the joint estimation of target and concatenation cost weights, we use only pairs of attached syllables ($s, s + 1$) in the held out data for which the first syllable s is detached from any preceding syllable. While searching for the N -best and N -worst candidates in the segment database, we now look for segment pairs which minimize the combined distance to the consecutive held-out syllables, s and $s + 1$. The sum of RMS distances for the pair of syllable segments are then set equal to the sum of the target costs of both syl-

lables plus the concatenation cost between the syllables, resulting in a system of linear equations. Note that because syllable s of the held-out pair is always detached, its target cost uses the independent weights, w_p^T , while syllable $s + 1$ uses the weights w_p^J and w_J which we are trying to estimate. In practice, we estimate w_p^T first using detached held-out segments as described in the previous section. These weights can then be plugged into the system of equations for the attached syllables, allowing the $P + 1$ unknown weights for attached syllables to be estimated using least-squares.

The weights for attached syllables are listed in Table 5. Most categorical features other than lexical stress are assigned zero weight due to the general dominance of the concatenation cost w_J . This is reasonable since, physiologically, segments within the same intonation phrase can not manifest sudden divergences from their continuous path. The attached syllable cost, therefore, becomes a trade-off between input F0 cost, T_{f0} , concatenation cost, J , and a lexical stress match, T_{lex} . For surprise and sadness, higher values of concatenation cost weight indicate the importance of voiced segment continuity in these emotions. Interestingly, for anger the subcost T_{f0} still plays an important role, as evidenced by its higher weight relative to the other emotions (0.68). For angry segments with similar costs, the segment with a more similar neutral counterpart in the corpus will be chosen at the expense of introducing small discontinuities.

4.3. Pruning

Even though Viterbi search is relatively efficient, the number of potential candidate units for each syllable is equal to the entire syllable corpus. Computation can be reduced significantly by pruning F0 segments that are unlikely given the input specification. To achieve this, we use a syllable duration criterion to eliminate contour segments with durations significantly different from the duration of the input. To do this we set a duration pruning range which is one tenth of the length of the input F0 segment. Hence, for example, if an F0 segment is 300 ms, the range is ± 30 ms, which results in pruning of all contours shorter than 270 ms and longer than 330 ms. Note that these thresholds assume that duration conversion is applied

Table 4
Estimated weights for detached syllables across three target emotions

| | Surprised | Sad | Angry |
|-------------|-----------|-------|-------|
| w_{lex} | 13.67 | 12.30 | 18.74 |
| w_{wpos} | 24.52 | 11.29 | 18.47 |
| w_{spos} | 11.33 | 4.91 | 3.31 |
| w_{pofs} | 1.13 | 4.82 | 8.82 |
| w_{ppofs} | 24.27 | 6.49 | 10.54 |
| w_{onset} | 15.08 | 0.33 | 5.54 |
| w_{coda} | 8.23 | 6.09 | 6.36 |
| w_{F0} | 0.47 | 0.69 | 1.00 |

Table 5
Estimated weights for attached syllables across three target emotions

| | Surprised | Sad | Angry |
|-------------|-----------|------|-------|
| w_{lex} | 17.89 | 6.43 | 15.98 |
| w_{wpos} | 0.0 | 0.0 | 0.0 |
| w_{spos} | 0.0 | 0.0 | 0.0 |
| w_{pofs} | 0.0 | 0.0 | 0.0 |
| w_{ppofs} | 0.0 | 0.0 | 0.0 |
| w_{onset} | 3.23 | 0.0 | 0.0 |
| w_{coda} | 0.0 | 0.0 | 8.74 |
| w_{F0} | 0.27 | 0.37 | 0.68 |
| w_J | 0.74 | 0.70 | 0.48 |

before F0 segment selection so that the duration pruning does not cause search errors when an emotion is characterized by markedly different durations compared to the neutral case.

5. Duration conversion

Neutral phone durations for vowels, nasals, glides and fricatives are transformed using a set of regression trees. The durations of all other broad classes are left unmodified. In building the regression trees, phone, syllable and word level contextual factors are used as categorical predictors as well as the continuous input duration (*dur*). The leaf nodes of the trees are trained to predict scaling factors rather than absolute durations, i.e. deviations relative to neutral speech are modeled rather than the absolute durations of the target emotion. In addition to the syllable and word level features listed in Section 2 (lexical stress, position in word, position in sentence, part of speech), features relating to the basic phonetic context including phone identity (*ph*), identity of the previous phone (*prev*) and identity of the next phone (*next*), are also included in the pool of regression tree features. The phone set consists of 16 vowels, 2 nasals, 4 glides and 9 fricatives which make up the phone identity values. Neighboring phone identity is expressed in terms of broad classes. The Matlab Statistics Toolbox implementation of classification and regression trees was used to build the trees. A minimum leaf occupancy count of 10 samples was set as a stopping condition while growing the trees. Trees were then pruned using 10-fold-cross-validation on the training data. The pruning level which minimized the prediction cost on the entire held out set was chosen for the final tree.⁴

During conversion, the sequence of phones in the test utterance and their durations are extracted along with the relevant contextual factors. For the experiments described below, the input durations are taken directly from the neutral utterances of the speaker. Each phone duration and context are then input into the appropriate broad class regression tree to generate a duration tier for the utterance. This duration tier is thus essentially a sequence of scaling factors which determine how much each phone in the utterance is going to be stretched or collapsed.

Trees were built using different features groups in order to select the best feature combination for each emotion and broad class based on RMS error (RMSE) between the predicted and target durations in the test data. RMS error is frequently used in the speech synthesis literature to evaluate duration prediction methods (see Goubanova and King, 2003; Iida et al., 2003). The contextual feature pool was grown by adding one or two new features at a time. The feature groups (FG) are listed in Table 6 and the best feature groups per emotion and broad class are listed in Table 7.

Table 6

The feature Groups tested for relative duration prediction

| | |
|-----------------------|-----------------------|
| Feature Group 0 (FG0) | Input duration |
| Feature Group 1 (FG1) | FG0 + phoneID |
| Feature Group 2 (FG2) | FG1 + leftID, rightID |
| Feature Group 3 (FG3) | FG2 + lex |
| Feature Group 4 (FG4) | FG3 + spos |
| Feature Group 5 (FG5) | FG4 + wpos |
| Feature Group 6 (FG6) | FG5 + pofs |

In general the RMSE values did not improve beyond the 0.025–0.035 s range. For glides and nasals the same feature combination, consisting of phone-level context and input duration, produced the minimum error across all emotions. Addition of higher level context did not improve the prediction of nasal and glide durations. For sadness, vowel and fricative durations also followed this pattern, where higher level context did not improve the RMS values. Vowels for anger also relied heavily on the neutral input durations, while fricative durations were best approximated using lexical stress in addition to input duration and phonetic context. For surprise, on the other hand, target vowel durations were better approximated using the higher level features lexical stress, position in word and position in sentence. Fig. 5 illustrates the tree used to convert neutral vowels to surprise. It is clear, for instance, that the vowel scaling factors are heavily dependent on whether the vowel is at the end of the sentence (i.e. in the last word) since this is the question at the root of the tree. Fricative durations for surprise also improved with the addition of lexical stress and position in sentence. This suggests that the duration effects of surprise are less constant and more context-sensitive than that of anger and sadness. Such a result is analogous to our findings in the F0 segment selection section, where position in sentence also gained a higher weight for surprise compared to other emotions.

6. Spectral conversion

A GMM-based spectral conversion method based on (Ye, 2005) is used to map each neutral spectrum to that of a desired target emotion. Line spectral frequencies (LSF) were used as the acoustic features to be converted. To train the conversion function, LSF parameter vectors of order 30 were computed for parallel pairs of neutral-emotional utterances. These were then time-aligned using the state-based phonetic alignments computed using HTK. The number of mixture components was set to 16. An Overlap and Add (OLA) synthesis scheme was used to combine the converted spectral envelope with the neutral (unmodified) residual. Fig. 6 illustrates the average spectra of all instances of vowel /ae/ in neutral, emotional, and converted test utterances in the case of sadness and anger. The average spectra of the vowel in converted utterances approximate the target emotion much better than the input neutral spectra. In general, the spectral conversion module produced a breathy voice quality for sadness as

⁴ Cross validation to find the best pruning level is a method recommended in the Matlab 7.0 Statistics Toolbox.

Table 7
Feature Groups (FG) which resulted in minimum RMS errors (RMSE) for all broad phone classes. RMSE is given in seconds

| | Vowels | | Glides | | Nasals | | Fricatives | |
|-----------|--------|-----|--------|-----|--------|-----|------------|-----|
| | RMSE | FG | RMSE | FG | RMSE | FG | RMSE | FG |
| Surprised | 0.0345 | FG5 | 0.0289 | FG2 | 0.0285 | FG2 | 0.0345 | FG4 |
| Sad | 0.0297 | FG2 | 0.0295 | FG2 | 0.0211 | FG2 | 0.0287 | FG2 |
| Angry | 0.0370 | FG1 | 0.0298 | FG2 | 0.0275 | FG2 | 0.0328 | FG3 |

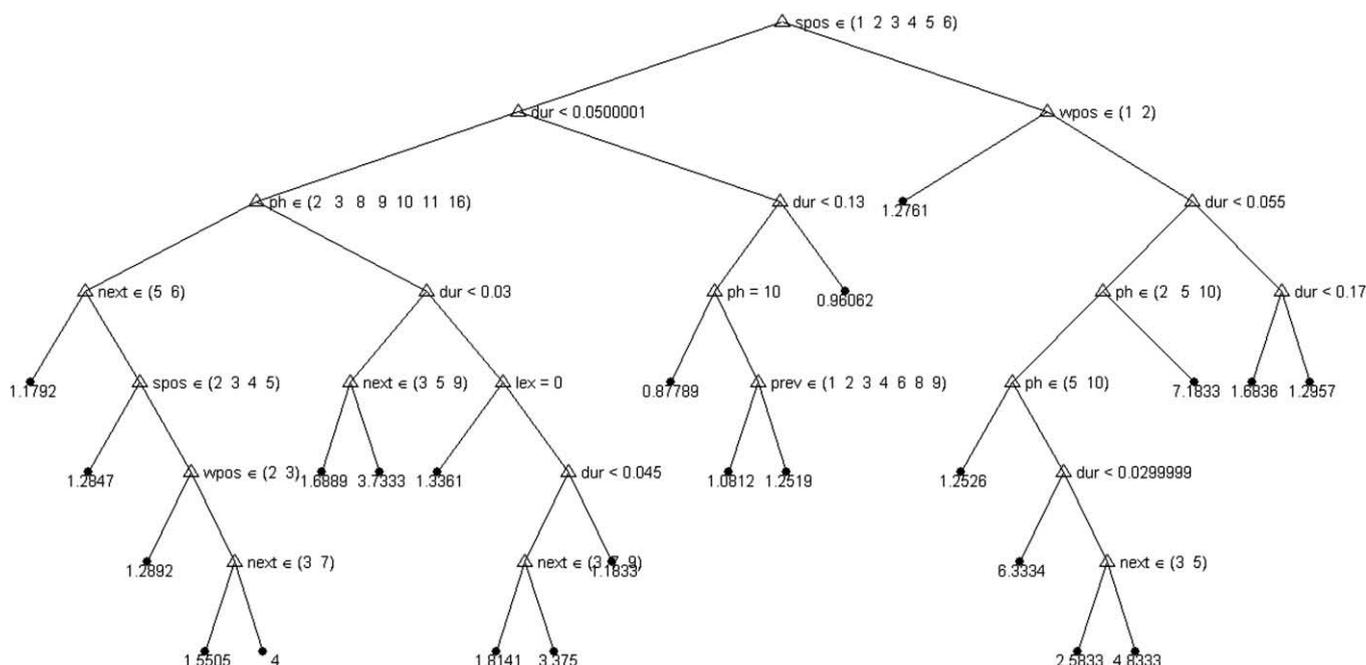


Fig. 5. Regression tree for converting vowel durations from neutral to surprised: *spos* refers to position in sentence, *wpos*, to position in word, *dur*, to input neutral duration in seconds, *ph* to phone identity, *prev* and *next*, to the broad class identities of the left and right context.

evidenced by a sharp spectral tilt in Fig. 6a and a tense voice quality for anger. The converted spectra for surprise also sounded slightly tense compared to the neutral input, although this tension did not necessarily make the utterance more surprised.

7. Perceptual evaluation

In order to evaluate each conversion module in isolation and integrated as a complete system, a series of perceptual listening tests were conducted using paid subjects who were asked to judge various properties of the converted utterances.⁵

7.1. Evaluation of spectral conversion

A preference test was conducted to evaluate the effect of spectral conversion on emotion perception. Subjects were asked to listen to versions of the same utterance and decide which one conveyed a given emotion more clearly. One ver-

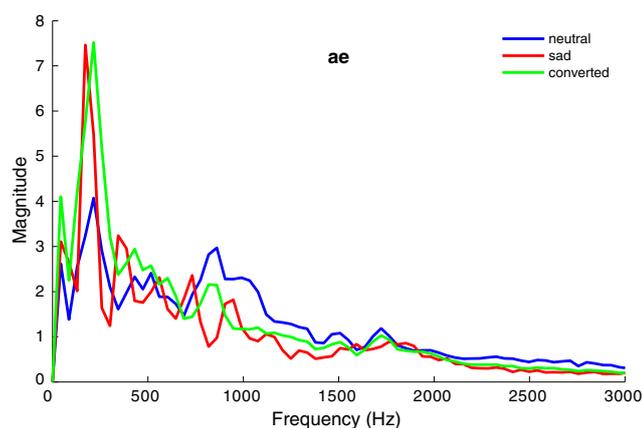
sion had spectral conversion applied while the other had the unmodified neutral spectrum. F0 contours for both utterances were identical and were generated for the target emotion by using the F0 segment selection method. No duration modification was applied for this test.

Twenty subjects participated in the evaluation. Each subject performed 15 comparisons, 5 in each emotion, resulting in 100 comparisons per emotion. The layout of the test for one emotion is illustrated in Fig. 7. The sample test utterances in each emotion were changed after the first ten subjects, in order to evaluate a wider range of utterances. Fig. 8 displays the percentage preference rates. As can be seen, spectral conversions were consistently preferred for anger and sadness (*t*-test, $p \ll 0.01$), while for surprise most people preferred unmodified spectra since the conversion did not seem to add a notable surprise element to the utterance and the original had a slightly crisper quality due to the lack of spectral processing.

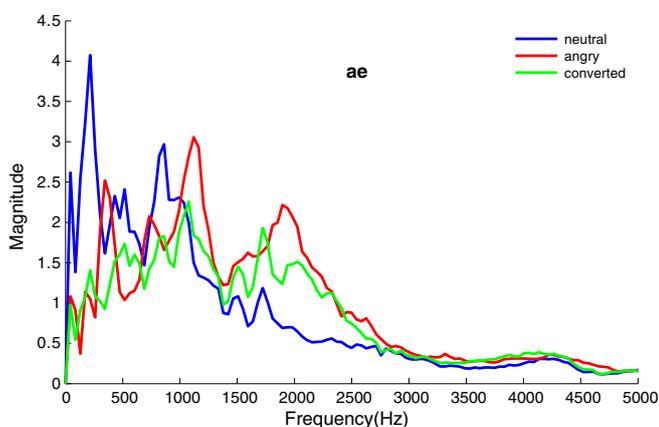
7.2. Evaluation of HMM-based F0 generation

The syllable-based HMM F0 generation was first compared with the baseline Gaussian normalization scheme

⁵ Speech samples output by the conversion system are available online at <http://mi.eng.cam.ac.uk/~zi201/conversions.html>.



(a) Average spectra of vowel /ae/ for neutral, sad and converted utterances.



(b) Average spectra of vowel /ae/ for neutral, sad and converted utterances.

Fig. 6. Long-term average magnitude spectra of vowel /ae/ taken from neutral, emotional and converted test utterances in the case of sadness and anger. (a) Average spectra of vowel /ae/ for neutral, sad and converted utterances. (b) Average spectra of vowel /ae/ for neutral, sad and converted utterances.

defined by Eq. (1) of Section 1. This baseline only takes advantage of the means and variances of the source and target expressive styles and hence relies heavily on the shape of the input neutral F0 contour. In order to show that the HMM-models driven by linguistic features outperform contours generated by this baseline, a preference test was conducted asking subjects to compare two utterances which were identical except for their pitch contours: in one of the utterances, the original pitch contour was converted using Gaussian normalization, and in the other it was generated by the syllable HMMs. The original neutral durations were left unmodified. For both utterances, spectral conversion to the target emotion was applied. As we have noted in the introduction, it is important to evaluate F0 generation jointly with spectral conversion since our informal tests showed that the F0 contours convey the emotion more strongly when combined with the appropriate voice quality.

Thirty subjects (15 male and 15 female) participated in this test. Twenty-one of the subjects were native speakers and of the remaining nine, English was a second language. Once again, they were asked to choose which one of the utterances they thought was angrier/more surprised/sadder. For each pair, the different F0 conversion methods appeared in random order. Five comparisons per emotion were presented to all subjects, resulting in 150 comparisons per emotion. The utterances were changed for every ten subjects to cover a wider range of sentences and contexts in the test set. This resulted in the evaluation of 15 unique sentences per emotion, each of which were evaluated by 10 subjects (Fig. 9). Overall, the subjects strongly preferred the HMM-generated contours for surprise (t -test, $p \ll 0.01$). This confirms that simply scaling neutral F0 segments does not really help convey the emotion and that actual segment shapes are better modeled using the HMMs. For anger, on the other hand, the overall preference scores did not point as strongly to one or the other method but the result was still significant ($p = 0.027$). In the case of sadness, HMM-based contours were preferred 67.3% of the time ($p \ll 0.01$). After completing the listening test, subjects were asked to write down the emotion they found easiest to choose between the options and the one they thought was the hardest. The surveys revealed that subjects were divided evenly between anger and sadness as the emotion for which they had most difficulty making a choice.

Which one sounds more surprised?

Please listen to each pair of utterances and choose the one which is able to express the target emotion (surprised) better.

| | |
|------------------------------|------------------------------|
| <input type="radio"/> A1.wav | <input type="radio"/> B1.wav |
| <input type="radio"/> A2.wav | <input type="radio"/> B2.wav |
| <input type="radio"/> A3.wav | <input type="radio"/> B3.wav |
| <input type="radio"/> A4.wav | <input type="radio"/> B4.wav |
| <input type="radio"/> A5.wav | <input type="radio"/> B5.wav |

Fig. 7. The layout of the preference test. Users are requested to choose the example in each pair which is most able to express the target emotion.

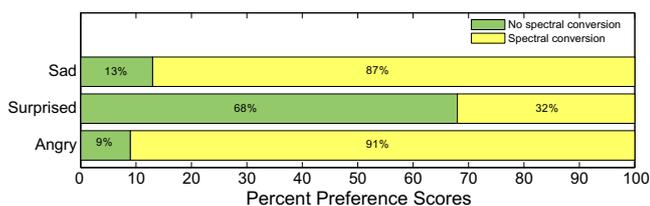


Fig. 8. Preference scores for GMM-based spectral conversion.

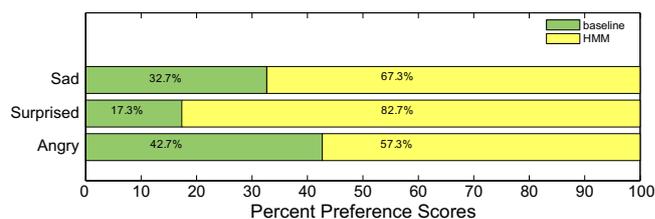


Fig. 9. Percent preference scores for syllable HMMs and Gaussian normalization (baseline).

7.3. Evaluation of segment selection

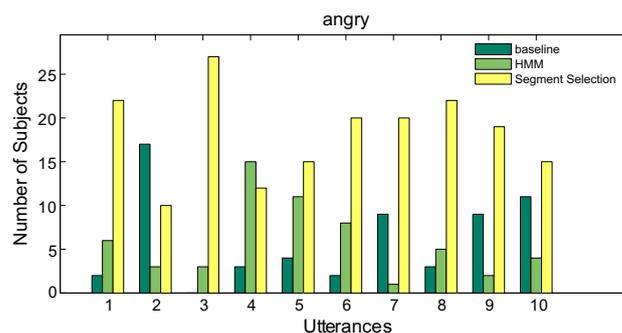
A three-way preference test was conducted in order to compare the F0 segment selection approach with the two methods evaluated in the previous section. Subjects were asked to compare three utterances which were identical except for the method used to convert the F0 contours: utterances converted using segment selection, syllable HMMs and Gaussian normalization were presented in random order. Spectral conversion was applied to all utterances but neutral durations were left unmodified. Thirty subjects participated in the test and each subject performed 10 comparisons per emotion. A total of 900 (30 × 10 × 3) comparisons were performed. The percentage preference scores per emotion are displayed as a stacked bar graph in Fig. 10 and the *p*-values resulting from *t*-tests for each pair of methods are shown in Table 8. As can be seen, for anger, segment selection was preferred significantly more frequently compared to the other methods. Unlike the previous test, however, the difference between the baseline and HMM-based contours was not significant (*p* = 0.083) in the case of anger. Segment selection was also significantly more popular when compared with the other two methods in the case of surprise (*p* << 0.01 in both cases). HMM-based contours were also still significantly more popular than those favoring the baseline. For sadness, HMM-based F0 generation was preferred half the time and the other half of subject preferences were split between the baseline and segment selection. There was however a significant tendency for segment selection when compared with the baseline (*p* = 0.02). The overall shift in preferences from HMM-based contours to those generated by segment selection, can be explained by the fact that the stored contour segments capture more realistic F0 movements in contrast to the HMM-generated contours which may be over-smoothed. Additionally, the incorporation of the input contour into the target cost function for segment selection may help select more appropriate segments (this argument is also supported by the high weight attached to this subcost for anger). Further information on this experiment including raw preference counts per speaker and per utterance can be found in (Inanoglu, 2008).

The distribution of preferences across each of the ten comparisons are illustrated in Fig. 11a–c. The segment selection method was strongly preferred for all conversions to anger except utterance 2 and utterance 4. In the utter-

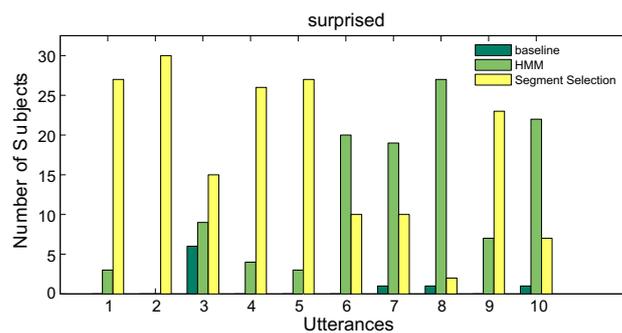
Table 8

The *p*-values resulting from *t*-tests performed on preferences for each pair of methods in the three-way preference test. Baseline, HMM and SegSel are used as abbreviations for Gaussian normalization, HMM-based F0 generation and segment selection respectively

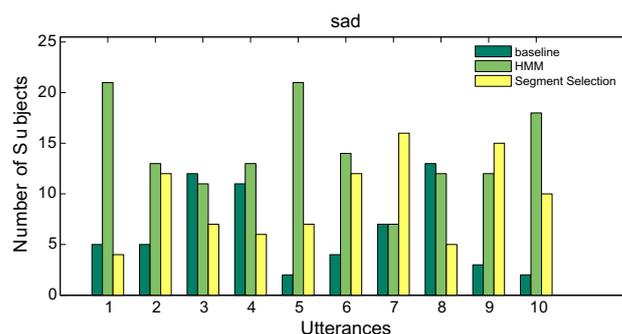
| | Baseline, HMM | Baseline, SegSel | HMM, SegSel |
|-----------|-----------------------|-----------------------|-----------------------|
| Angry | 0.083 | 9.4×10^{-16} | 4.8×10^{-16} |
| Surprised | 1.8×10^{-19} | 1.3×10^{-28} | 4.8×10^{-8} |
| Sad | 2.3×10^{-7} | 0.02 | 6.1×10^{-4} |



(a) Angry



(b) Surprised



(c) Sad

Fig. 11. Utterance specific analysis of preferences across three methods of F0 conversion.

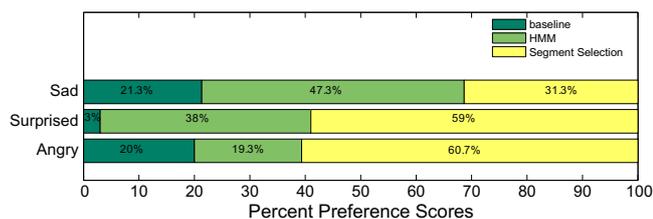


Fig. 10. Preference scores for each method and emotion.

ance-specific analysis of surprise (Fig. 11b), it may be observed that the segment selection method is not consistently preferred as in the case of anger. In fact, there are some utterances where subjects strongly prefer the HMM-based method and there are others where segment-selection is clearly preferred, which suggests that both

methods can be effective for surprise. The analysis of sadness across utterances is not as straightforward, since all methods generate convincingly sad sounding contours particularly when combined with the breathy voice quality which results from spectral conversion. Overall, the HMM-based method was selected most frequently but otherwise there was little consistency in the results. These scores suggest that most subjects were able to reduce their choices down to two and then had to guess which one of the remaining two is sadder. In fact, when subjects were asked explicitly which emotion they had most difficulty choosing, 70% recorded a difficulty with sadness compared to 40% reported in the two-way test of the previous section (Table 9). With the introduction of the segment selection approach, the difficulty with anger seems to have been resolved since only 13% of the subjects listed it as the emotion they had difficulty with compared with 43.3% from the previous section. Surprise continued to be an easy emotion to identify even with the two competing methods of HMMs and segment selection.

7.4. Evaluation of duration conversion

In the previous two tests, the focus was on evaluating F0 contours generated by different methods leaving the neutral durations unmodified. In this section, the contribution of duration conversion to the perception of a target emotion is evaluated. The test organization was similar to that shown in Fig. 7. Each subject had to listen to two utterances and decide which one sounded angrier/sadder/more surprised. Both utterances had their spectra converted. In one utterance, neutral phone durations were modified using the scaling factors predicted by the relative regression trees. In the other, they were left unmodified. Additionally, segment selection was applied to both utterances to replace the neutral pitch contours. Note that the F0 segments selected by this method actually depend on the input durations. Therefore, for some utterance pairs, the F0 contours were not identical, i.e. the contours that are appropriate for the modified durations may be different from those selected for the neutral syllable durations. The preference test therefore evaluated the joint effects of duration conversion and segment selection relative to the no duration conversion case.

The same 30 subjects participated in the test, where each subject performed 10 comparisons per emotion (Fig. 12). The results of the tests showed that converted durations

Table 9
Percent of subjects who identify a given emotion as “hardest to choose” in the two-way test described in Section 8.2 and the three-way test described in this section

| | Two-way test (%) | Three-way test (%) |
|-----------|------------------|--------------------|
| Angry | 43.3 | 13.3 |
| Surprised | 16.7 | 16.7 |
| Sad | 40 | 70 |

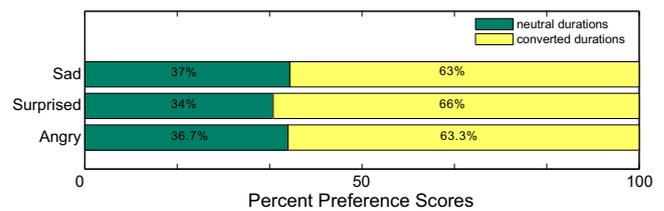
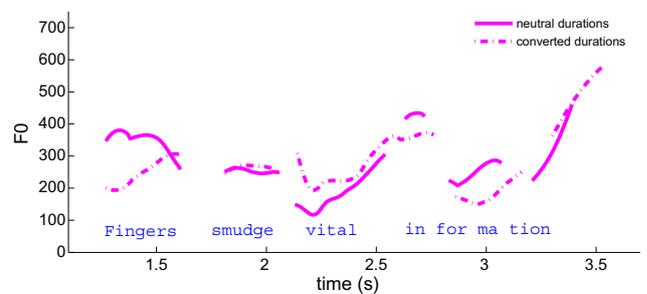


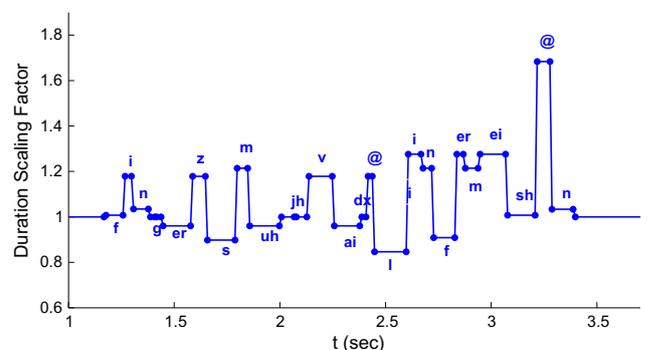
Fig. 12. Preference scores for duration conversion in each emotion.

were preferred more frequently than unmodified durations. This was very significant for all emotions ($p \ll 0.01$). The preferences for converted durations were slightly stronger for the case of surprise. In fact, none of the subjects listed “surprise” as an emotion they had difficulty with.

Fig. 13a illustrates an example of a surprised utterance where duration conversion was preferred strongly. Both F0 contours resulting from the different phone durations were plotted for comparison. The corresponding duration tier is also included in Fig. 13b, where the scaling factors for each neutral phone are identified explicitly. The horizontal line indicates a scaling factor of 1 i.e. no change. Even though the overall contour shape does not change very dramatically for the two cases, the durations and intonation of the final word “information” conveys surprise much more effectively with the scaled durations. All nasals and vowels are stretched in this final word, which results in the selection of a lower pitch movement in the lexically stressed syllable “/m/-/ei/”. Furthermore, the duration of the vowel /@/ in the final unstressed syllable is almost dou-



(a) F0 before and after duration conversion



(b) Corresponding duration tier

Fig. 13. Results of F0 segment selection for utterance “Fingers smudge vital information” before and after duration conversion (a) and the corresponding duration tier (b).

bled providing the time necessary for the final rise to reach a higher target. The combination of a very low stressed syllable with a gradual high rise, results in a question-like intonation that sounds amazed/surprised. The durations themselves provide room for this expression, and indirectly control the F0 movements selected by the search algorithm.

Contrary to this example, there were two of the ten utterances where subjects did not consistently prefer duration conversion. This is thought to occur when the sequence of neutral durations are already quite likely in the target emotion. In such cases, further modification of durations can be ineffective. Overall, however, duration conversion will often improve emotion conversion and rarely impair it. We therefore conclude that it is better to include it consistently in a conversion system framework.

7.5. Overall emotion classification performance

A final evaluation of the full conversion system was performed using a multiple-choice emotion classification test, where subjects were asked to identify the emotion in an utterance. To avoid forcing the subjects to choose an emotion when they were unsure, a “Can’t decide” option was included in the available choices.

To provide a basis for comparison, we first report the results of this test using the original natural utterances of the voice talent used to record the database. In practice, this test was conducted after the test on converted utterances in order to avoid a bias towards the original speech. Five utterances per emotion were presented to 30 subjects and the confusion matrix for listener preferences is summarized in Table 10. The results show that anger and sadness were communicated very clearly by the speaker, while there was some confusion with anger in the perception of surprise.

The results of the emotion classification test using converted neutral utterances generated by our conversion system are reported in Tables 11 and 12. Ten utterances per emotion were classified by 30 subjects in random order. Duration conversion and spectral conversion were applied to all outputs. Additionally, there were two hidden groups within each emotion: five of the conversions were synthesized using HMM-based contours and the other five were synthesized using segment selection. Confusions between emotions were analyzed separately for the two F0 conversion methods.

The conversion outputs using HMM-based F0 contours conveyed sadness as well as the original sad speech, while

Table 10
Percent confusion scores for the emotion classification task of original emotional utterances spoken by the female speaker

| | Angry (%) | Surprised (%) | Sad (%) | Cannot decide (%) |
|-----------|-----------|---------------|---------|-------------------|
| Angry | 99.3 | 0.7 | 0 | 0 |
| Surprised | 20.0 | 66.0 | 0 | 14.0 |
| Sad | 0.7 | 0 | 96.0 | 3.3 |

Table 11

Percent confusion scores for the emotion classification task for utterances where HMM-based contours are used

| | Angry (%) | Surprised (%) | Sad (%) | Cannot decide (%) |
|-----------|-----------|---------------|---------|-------------------|
| Angry | 64.7 | 8.0 | 4.7 | 22.6 |
| Surprised | 10.0 | 60.7 | 0 | 29.3 |
| Sad | 0.7 | 0.7 | 96.0 | 2.6 |

Table 12

Confusion scores for the emotion classification task for utterances where F0 segment selection is used

| | Angry (%) | Surprised (%) | Sad (%) | Cannot decide (%) |
|-----------|-----------|---------------|---------|-------------------|
| Angry | 86.7 | 0.7 | 0 | 12.6 |
| Surprised | 8.7 | 76.7 | 0 | 14.7 |
| Sad | 0.7 | 0 | 87.3 | 12 |

the recognition rate for surprise (60.7%) was slightly lower than that of the naturally spoken surprised speech (66%) and the rate for anger (64.7%) was much lower than that of the naturally spoken anger (99.3%). There was considerable indecision amongst subjects when classifying surprise and anger. Overall there was moderate inter-rater agreement as given by Fleiss’ kappa statistic ($\kappa = 0.506$).

With segment selection, the classification rate for anger increased significantly up to 86.7%. This indicates that appropriate F0 prediction is a critical component of anger despite the fact that it is normally considered to be an emotion with dominant spectral characteristics. Surprise is also recognized better using segment selection (76.7%), indeed, the converted surprise utterances were identified more accurately than the naturally spoken surprised utterances. This may be explained by the spectral conversion module which tends to over-smooth the converted spectra slightly. In the naturally spoken utterances, there is a tension in some of the surprised speech which may have created confusion between anger and surprise. This tension is reduced and more consistent in the converted surprised speech. The same effect, however, may have slightly reduced the recognition rates for anger, since the smoothing in that case resulted in conversions which did not sound as harsh as the naturally spoken angry utterances. Overall, the effect of F0 prediction method on emotion recognition rates was significant for all emotions. Segment selection resulted in better recognition in the case of anger ($p = 0.0006$) and surprise ($p = 0.004$), while HMM-based contours resulted in higher recognition scores for sadness ($p = 0.018$). The inter-rater agreement was also much higher with segment selection ($\kappa = 0.635$).

Finally, as part of the emotion classification test, we also asked subjects to categorize each utterance in terms of intonation quality using the options “Sounds OK” or “Sounds Strange.” The intonation quality ratings are illustrated in bar charts for each method (Figs. 14 and 15). The effect of F0 prediction method on quality was significant only in the case of surprise ($p = 0.0006$). For both methods,

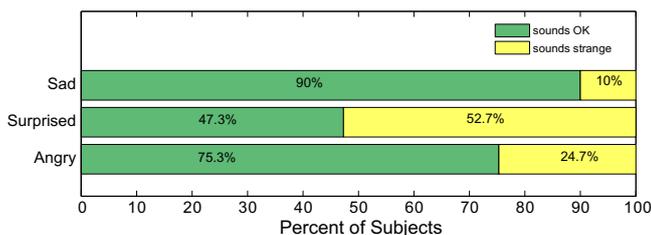


Fig. 14. Categorical quality ratings for spectral conversion + duration conversion + HMM-based contour generation.

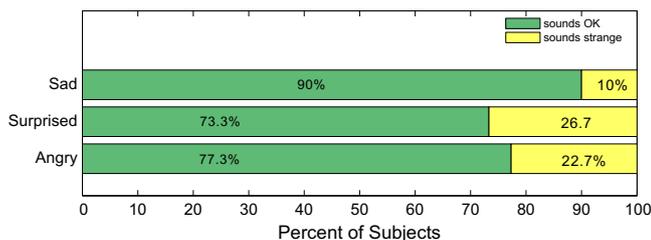


Fig. 15. Categorical quality ratings for spectral conversion + duration conversion + F0 segment selection.

the percentage quality ratings for sadness are identical and generally very high (90% “sounds OK”). Subjects also thought that both methods attempted to convey anger naturally most of the time, even though the actual emotion recognition rates are very different between the methods. For surprise, on the other hand, quality perception improved significantly with segment selection, where 73.3% of conversions sounded OK compared with only 47.3% when HMM-based contours were used. Therefore, unlike anger, the recognition rates and quality ratings for surprise were somewhat correlated.

8. Conclusions

A system for emotion conversion in English has been described which consists of a cascade of modules for transforming F0, durations and short-term spectra. The system was evaluated using three target emotions for which emotional speech data was collected. In principle, the system should be scalable to any emotion which modifies the acoustic-prosodic characteristics of speech. Two different syllable-based F0 conversion techniques were implemented and evaluated as well as a duration conversion method which performs transformation on the segmental level. Subjective preference tests confirmed that each module augments emotional intensity when combined with the others. The full conversion system with either F0 prediction method was able to convey the target emotions above chance level. However, F0 segment selection produced more natural and convincing expressive intonation compared to syllable HMMs, particularly in the case of surprise and anger.

The different modules also indirectly reveal interesting characteristics of the target emotions. For example, examining the weights in the case of segment selection highlight

the contextual factors which have a more dominant role in the expression of each target emotion. In general, surprise was found to be an emotion which is highly dependent on syllable and word-level linguistic factors. On the other hand, the F0 and duration characteristics of angry speech relied heavily on the information in the input F0 contours and durations and less so on the contextual factors.

Finally, using only a modest amount of training data, the perceptual accuracy achieved by the complete conversion system was shown to be comparable to that obtained by a professional voice talent. Hence it may be concluded that the conversion modules which have been described in this paper provide an effective and efficient means of extending a single emotion TTS system to exhibit a range of expressive styles. Furthermore, the conversion system provides a flexible framework for investigating alternative cost functions for segment selection or alternative dissimilarity measures for duration and F0 prediction. Such future investigations can be compared to the current system to quantify perceptual improvements.

References

- Banziger, T., Scherer, K., 2005. The Role of Intonation In Emotional Expressions. *Speech Comm.* 46, 252–267.
- Barra, R., Montero, J., Arriola, J.G., Guaras, J., Ferreiros, J., Pardo, J., 2007. On the limitations of voice conversion techniques in emotion identification. In: *Proc. Interspeech*.
- Boersma, P., Weenink, D., 2005. Praat: doing phonetics by computer. <<http://www.praat.org>>.
- Bulut, M., Lee, S., Narayanan, S., 2007. A statistical approach for modeling prosody features using POS tags for emotional speech synthesis. In: *Proc. ICASSP*.
- Fallside, F., Ljolje, A., 1987. Recognition of isolated prosodic patterns using hidden Markov models. *Speech Lang.* 2, 27–33.
- Gillett, B., King, S., 2003. Transforming F0 contours. In: *Proc. Eurospeech*.
- Goubanova, O., King, S., 2003. Using Bayesian belief networks to model duration in text-to-speech synthesis. In: *Proc. Internat. Congr. on Phonetic Sciences*, Vol. 3. p. 2349.
- Helander, E., Nurminen, J., 2007. A novel method for pitch prediction in voice conversion. In: *Proc. ICASSP*, Vol. 4. pp. 509–512.
- Iida, A., Campbell, N., Higuchi, F., Yasamura, M., 2003. A corpus-based speech synthesis system with emotion. *Speech Comm.* 40, 161–187.
- Inanoglu, Z., 2003. Pitch transformation in a voice conversion framework. Master's Thesis, University of Cambridge.
- Inanoglu, Z., 2008. Data-driven Parameter Generation for Emotional Speech Synthesis. Ph.D. Thesis, University of Cambridge.
- Inanoglu, Z., Young, S., 2005. Intonation modelling and adaptation for emotional prosody generation. In: *Proc. Affective Computing and Intelligent Interaction*.
- Jensen, U., Moore, R., Dalsgaard, P., Lindberg, B., 1994. Modelling intonation contours at the phrase level using continuous density hidden Markov models. *Comput. Speech Lang.* 8, 227–260.
- Jones, D., 1996. *Cambridge English Pronunciation Dictionary*. Cambridge University Press.
- Kain, A., Macon, M., 1998. Spectral voice conversion for text-to-speech synthesis. In: *Proc. ICASSP*, Vol. 1. pp. 285–288.
- Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., Shikamo, K., 1999. GMM-based voice conversion applied to emotional speech synthesis. *IEEE Trans. Speech Audio Process.* 7 (6), 697–708.
- Ross, K., Ostendorf, M., 1994. A dynamical system model for generating F0 for Synthesis. In: *Proc. ESCA/IEEE Workshop on Speech Synthesis*.

- Scherer, K., Banziger, T., 2004. Emotional expression in prosody: a review and agenda for future research. In: *Proc. Speech Prosody*.
- Schroder, M., 1999. Emotional speech synthesis – a review. In: *Proc. Eurospeech*, Vol. 1. pp. 561–564.
- Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C., 1992. ToBI: a standard scheme for labelling prosody. In: *Proc. ICSLP*, Vol. 40. pp. 862–869.
- Stylianou, Y., Cappe, O., Moulines, E., 1998. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech Audio Process.* 6, 131–142.
- Tao, J., Yongguo, K., Li, A., 2006. Prosody conversion from neutral speech to emotional speech. *IEEE Trans. Audio Speech Lang. Process.* 14, 1145–1153.
- Tian, J., Nurminen, J., Kiss, I., 2007. Novel Eigenpitch-based prosody model for text-to-speech synthesis. In: *Proc. Interspeech*.
- Toda, T., Tokuda, K., 2005. Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. In: *Proc. Interspeech*.
- Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., Kitamura, T., 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In: *Proc. ICASSP*, Vol. 3. pp. 1315–1318.
- Tokuda, K., Zen, H., Black, A., 2002. An HMM-based speech synthesis system applied To English. In: *IEEE Speech Synthesis Workshop*.
- Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., 2002. Multi-space probability distribution HMM. *IEICE Trans. Inform. Systems* 3, 455–463.
- Tsuzuki, H., Zen, H., Tokuda, K., Kitamura, T., Bulut, M., Narayanan, S., 2004. Constructing emotional speech synthesizers with limited speech database. In: *Proc. ICSLP*, Vol. 2. pp. 1185–1188.
- Vroomen, J., Collier, R., Mozziconacci, S., 1993. Duration and intonation in emotional speech. In: *Proc. Eurospeech*, Vol. 1. pp. 577–580.
- Wu, C.H., Hsia, C.-C., Liu, T.-E., Wang, J.-F., 2006. Voice conversion using duration-embedded Bi-HMMs for expressive speech synthesis. *IEEE Trans. Audio Speech Lang. Process.* 14 (4), 1109–1116.
- Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., 2003. Modeling of various speaking styles and emotions for HMM-based speech synthesis. In: *Proc. EUROSPEECH*, Vol. 3. pp. 2461–2464.
- Ye, H. 2005. High-quality voice morphing. Ph.D. Thesis, Cambridge University.
- Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., 2004. An acoustic study of emotions expressed in speech. In: *Proc. ICSLP*.
- Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1998. Duration modelling for HMM-based speech synthesis. In: *Proc. ICSLP*.
- Young, S.J. et al., 2006. The HTK Book Version 3.4, Cambridge University. <<http://htk.eng.cam.ac.uk>>