HAL
open science

Data-driven Emotion Conversion In Spoken English

Zeynep Inanoglu, Steve Young

To cite this version:

Zeynep Inanoglu, Steve Young. Data-driven Emotion Conversion In Spoken English. Speech Communication, 2009, 51 (3), pp.268. 10.1016/j.specom.2008.09.006 . hal-00499236

HAL Id: hal-00499236

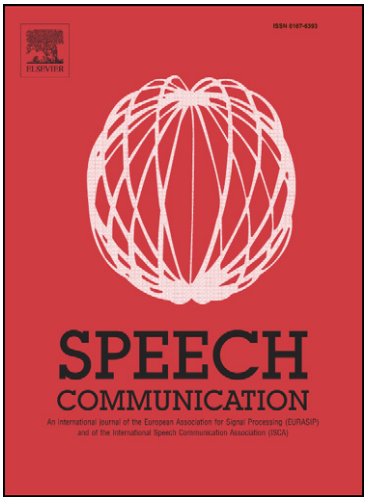
https://hal.science/hal-00499236

Submitted on 9 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Accepted Manuscript



Data-driven Emotion Conversion In Spoken English

Zeynep Inanoglu, Steve Young

PII: S0167-6393(08)00136-2
DOI: 10.1016/j.specom.2008.09.006
Reference: SPECOM 1752

To appear in: *Speech Communication*

Received Date: 29 May 2008
Revised Date: 9 September 2008
Accepted Date: 9 September 2008

Please cite this article as: Inanoglu, Z., Young, S., Data-driven Emotion Conversion In Spoken English, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.09.006

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Data-driven Emotion Conversion In Spoken English

Zeynep Inanoglu, Steve Young *

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

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 minutes) and which have been evaluated for three target emotions: anger, surprise and sadness. Two alternative F0 generation methods are presented. Firstly, an HMM-based approach uses linguistic features at the syllable level to model F0 segments in an effort to capture the affective and linguistic layers of intonation within a single framework. Secondly, a segment selection approach utilises a concatenative framework to directly search for F0 segments in the training corpus. In either case, phone durations are transformed using a set of regression trees and a GMM-based spectral conversion technique 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 emotion and an emotion classification test showed that both methods of F0 generation communicate 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. Furthermore, using segment selection, the emotion recognition rates for the converted neutral utterances were comparable to the same utterences spoken directly in the target emotion.

Key words: 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 [1] 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. However, whilst conceptually emotion conversion can be thought of as just another 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 F0 and duration patterns[2][3][4]. 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 such as GMM based F0 transfor-

* Corresponding author.
Email addresses: zeynep@gatesscholar.org (Zeynep Inanoglu), sjy@eng.cam.ac.uk (Steve Young).

ACCEPTED MANUSCRIPT

mation[5], piecewise linear transformation based on salient points in the contour[6] and codebook-based approaches used to predict entire F0 segments using linguistic information[7]. 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 [8][9][10], the focus of this paper will be on the transformation and evaluation of prosody in a unified emotion conversion framework. Similar to [7], 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 which attempts to analyze the interaction between part of speech tags, sentence position and emotional F0 contours support this modeling approach [11].

Various methods of emotion conversion have been reported in the literature. In [12], 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. In [14], a unified conversion system was proposed using duration embedded Bi-HMMs to convert neutral spectra and decision trees to transform syllable F0 segments. In [15], 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 [16][17].

In this paper we describe an emotion conversion system for English which is independent of the underlying synthesis system. It can therefore add an additional layer of expressiveness to an existing system without sacrificing quality. Prosody and voice quality are converted using methods which operate at different time intervals: F0 is modelled at the syllable level, duration is modelled at the phone level and voice quality is modelled at the speech frame level. The conversion system thus consists of the following modules:

(i) **F0 Generation**
   (a) **Syllable HMMs:** a generative method for modeling and synthesizing expressive F0 contours using syllable HMMs based on a small pool of linguistic features.
   (b) **Segment selection:** as in unit-based speech synthesis, this method expresses F0 conversion as a search problem, where actual syllable segments from the target emotion are spliced together under contextual and physiological constraints.

(ii) **Duration Conversion:** a set of regression trees specific to each broad phone class are used to scale neutral phone durations for a given target emotion.

(iii) **Spectral Conversion:** A GMM-based linear transformation method applied to the vocal tract spectrum in order to change vowel quality. The methods used are similar to the work of [12].

Each of the two alternative methods for F0 generation are compared within the full-conversion framework.

The rest of the paper is organised as follows. In section 2, the HMM-based F0 generation method is described and in section 3, the alternative F0 segment selection method is presented. In section 4, the duration conversion module is described and section 5 provides an overview of the spectral conversion module. The experimental setup of the conversion system is outlined in section 6. Finally, in section 7, the results of a perceptual study are reported.

2. F0 Generation From Syllable HMMs

HMMs have been used for the recognition of F0 related features such as accent and tone for some time [18][19]. 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, [20] 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 the syllable is widely considered to be the smallest temporal layer of F0 movement [21]. Hence, the system described here models F0 at the syllable level based on features derived from 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 [22]. Furthermore, because detailed information regarding the phonetic identities of syllabic constituents is ignored, the training data requirements for such a model set is inherently much smaller than that of a complete speech synthesis system.

2.1. *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 form, symbol or label which provides a link to F0 movements. Pitch accent classification schemes such as TOBI (Tones and Breaks Indices System) have been used to model and understand F0 movements in neutral speech [23], [24]. 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 emotional renderings of a given utterance. These features are lexical stress (lex), position in word (wpos), position in sentence (spos), part of speech of current word (pofs),

ACCEPTED MANUSCRIPT

Table 1
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%

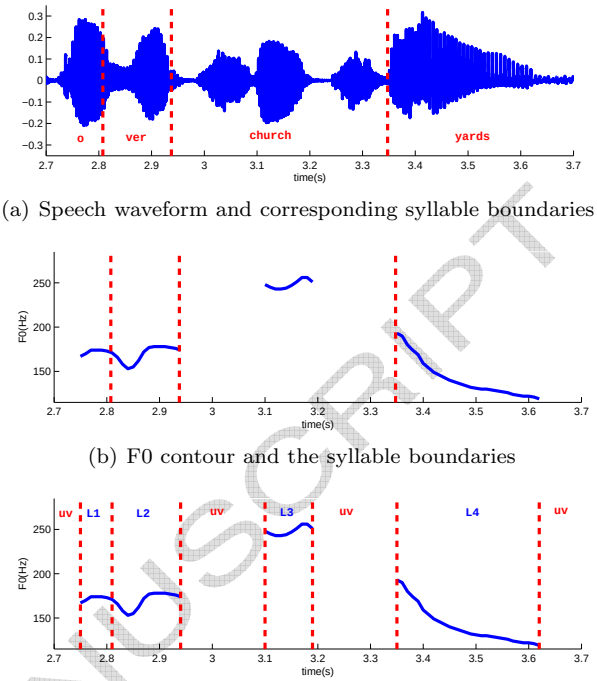
part of speech of previous word (ppofs), onset type (onset) and coda type (coda), where the onset and coda are either voiced, unvoiced or sonorant. Position in the sentence is identified explicitly for the first three words and the last three words in the sentence. All words in between these sentence-initial and sentence-final groups are identified with a single tag (spos=4). Syllable position in the word can take on one of 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 [1] .

Even though intonation movements are most meaningful at the syllable level, microprosodic effects can be observed at the segmental level: for instance, [25] finds strong effects from the consonant class on the following vowel. Such findings motivated the use of a broad classification scheme for the onset and coda of the syllable. 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 whilst keeping the context space as compact as possible [2] .

Although not all feature permutations are feasible, for example, there are syntactic limitations on part of speech tag sequences, the potential set of features is still large and hence some form of clustering is needed. To gain some insight into this, Table 1 summarizes the percentage of unseen feature combinations in a test set (see section 6.1 below) of 28 utterances given a training set of 272 utterances containing a total of 2086 unique features combinations. As can be seen, although 42.3% of the combinations in the test set are unseen, only matching 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 training data.

2.2. *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 used in HMM-based synthesis [26]. The voiced segment within each syllable was aligned with the context-dependent syllable models determined by the corresponding linguistic features. The unvoiced regions in the training utterances were modeled



(a) Speech waveform and corresponding syllable boundaries

(b) F0 contour and the syllable boundaries

(c) Syllable label boundaries after MSD-HMM training.

Fig. 1. Example syllable alignment for the phrase "over churchyards". Labels L1, L2, L3, and L4 represent linguistic feature combinations.

using a separate *uv* model which was always aligned with a zero-dimensional unvoiced symbol as defined in the MSD-HMM framework. Figure 1 illustrates an example of label alignments for a short speech segment.

The F0 model training follows a conventional HTK recipe[13]. The basic model topology is a three state left-to-right HMM 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. Separate 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

[1] The part of speech tags were generated using the proprietary tagger of Toshiba Research Speech Technology Group.
[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.

ACCEPTED MANUSCRIPT

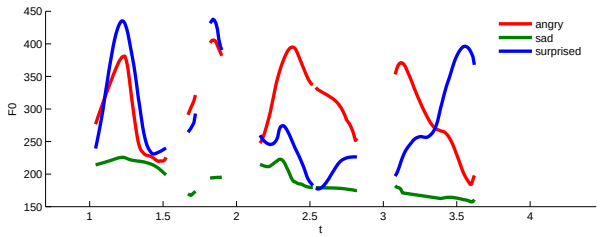


Fig. 2. HMM-generated contours for three emotions using the same utterance "Earwax affects koala bears terribly"

were clustered separately. The initial set of 6258 (2086 x 3) states are reduced to 801, 779 and 529 clusters for surprise, anger and sadness respectively.

2.3. *Generation From Syllable HMMs*

To generate an F0 contour, the required syllable label sequence is derived from the 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. Parameter generation used the HTS framework [3] in the mode where the state and mixture sequences are known [27], the latter being determined from the syllable boundaries and the duration models [28]. 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 Figure 2, 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 models. 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 synthesis. A method has recently been proposed 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) [29]. 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 estimation framework. When applied to our syllable F0 models, GV made a small difference to the overall utterance variances for surprise (Figure 3). However, for anger and sadness, the addition of global variance to

[3] HTS Version 2.1α was used

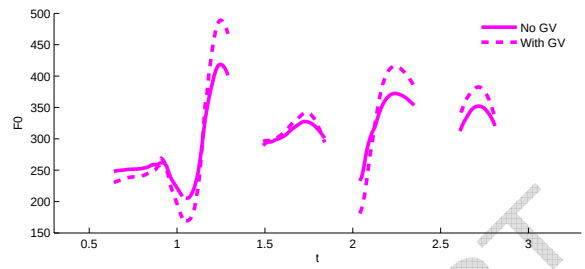


Fig. 3. Surprised F0 contour for the utterance "Dry black-thorn is grim stuff" with and without Global Variance (GV) based parameter generation

the parameter generation framework made no perceptual difference. Hence, in the perceptual evaluations of HMM-based F0 contours described below, GV-based parameter generation is only used in the case of surprise.

3. 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 to predict F0 contours in a non-expressive TTS framework from a large corpus of Mandarin speech [30]. The goal of the method described here, however, is to generate expressive prosody from limited data in a conversion framework. Parallel neutral and emotional syllable F0 segments are stored as part of the unit definition as well as their common linguistic context. The same linguistic features are used as for the HMM-based system described in section 2. 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 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

ACCEPTED MANUSCRIPT

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

where L is the length after interpolating the two segments to have equal duration.

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.

3.1. *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. 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 a $2NM \times P$ matrix of subcosts, W is the $P \times 1$ vector of unknown weights and D is the $2NM \times 1$ vector of 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. The weights obtained for detached syllables are listed in Table 2. 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.

Table 2
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

3.2. *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:

ACCEPTED MANUSCRIPT

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

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 syllables 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 a least-squares.

The weights for attached syllables are listed in Table 3. 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.

Table 3
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

3.3. *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 300ms, the range is $\pm 30ms$, which results in pruning of all contours shorter than 270ms and longer than 330ms. Note that these thresholds assume that duration conversion is applied 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.

4. 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 (origdur). 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. To avoid data sparsity, 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.

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 ut-

ACCEPTED MANUSCRIPT

terance is going to be stretched or collapsed.

Table 4
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

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. The feature pool was grown by adding one or two new features at a time. The feature groups (FG) are listed in Table 4 and the best feature groups per emotion and broad class are listed in Table 5.

In general the RMSE values did not improve beyond the 25-35ms 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. 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. Figure 4 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 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. For anger, simply using the input duration along with phone identity yielded the minimum error for vowel durations. Once again, anger seems to rely heavily on the patterns in the neutral input utterance. Fricative durations for anger were best approximated using lexical stress in addition to neutral duration and phonetic context.

Table 5
Feature Groups (FG) which resulted in minimum RMS errors (RMSE) for all broad phone classes. RMSE is given in milliseconds

	Vowels		**Glides**		**Nasals**		**Fricatives**	
	RMSE	FG	RMSE	FG	RMSE	FG	RMSE	FG
surprised	34.47	FG5	28.98	FG2	28.55	FG2	34.48	FG4
sad	29.69	FG2	29.48	FG2	21.09	FG2	28.67	FG2
angry	36.99	FG1	29.80	FG2	27.53	FG2	32.79	FG3

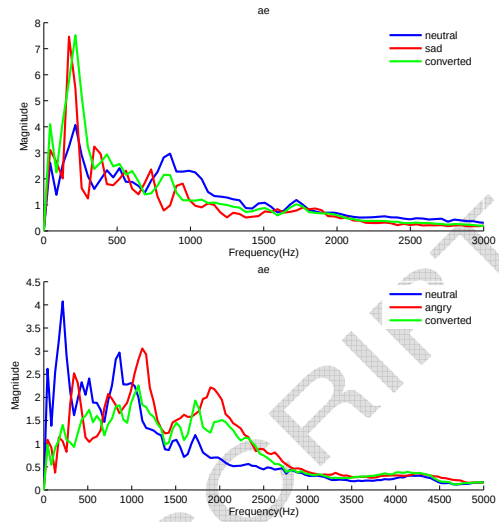


Fig. 5. Long-term average magnitude spectra of vowel /ae/ taken from neutral, emotional and converted test utterances in the case of sadness and anger.

5. Spectral Conversion

A GMM-based spectral conversion method is used to map each neutral spectrum to that of a desired target emotion [2][3]. 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. Figure 5 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 evidenced by a sharp spectral tilt and a harsh 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.

6. Experimental Setup

A block diagram of the complete emotion conversion system is illustrated in Figure 6. Spectral conversion is performed using pitch-synchronous LPC analysis/synthesis as the first step. Durations and F0 contours of the input utterance are then modified using the Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) implementation provided by the Praat software [32]. If duration conversion is performed, new syllable durations are com-

ACCEPTED MANUSCRIPT

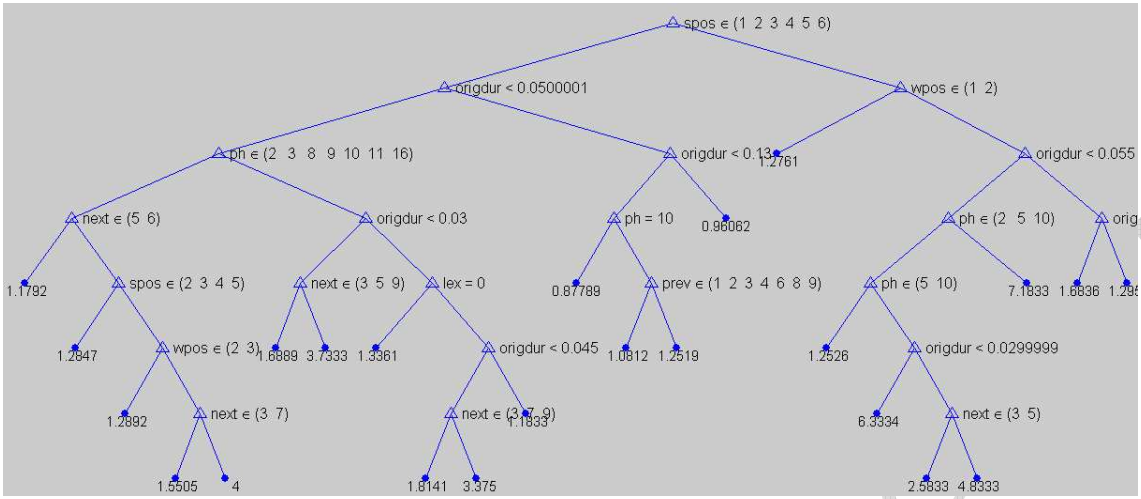


Fig. 4. Regression tree for converting vowel durations from neutral to surprised: *spos* refers to position in sentence, *wpos*, to position in word, *origdur*, to input neutral duration in seconds, *ph* to phone identity, *prev* and *next*, to the broad class identities of the left and right context. For a more detailed discussion of the trees, see [31]

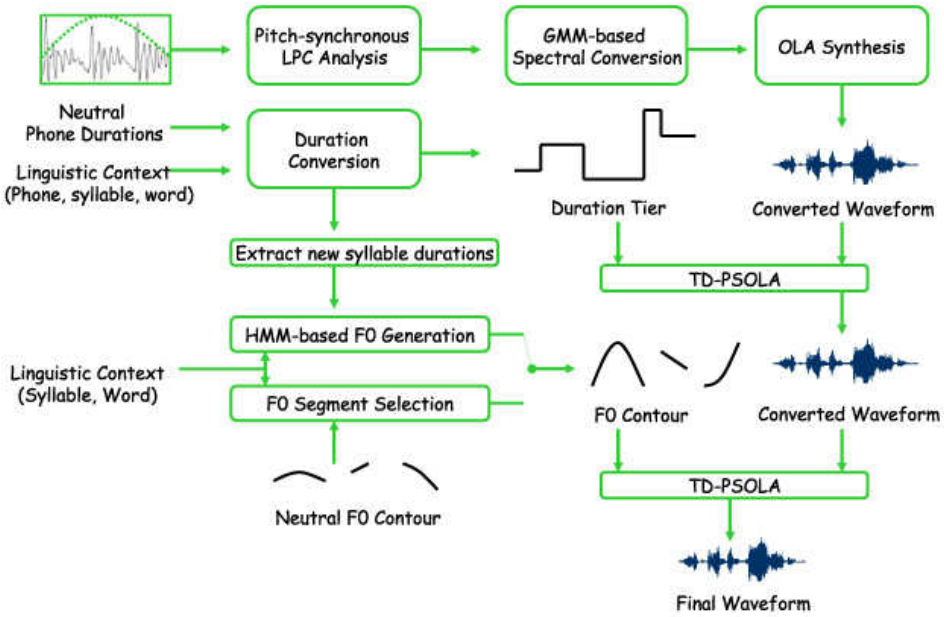


Fig. 6. Flowchart of Emotion Conversion System

puted and input into the selected F0 generation modules. The generated F0 contour and is then applied using a second pass of TD-PSOLA conversion. Speech samples output by the conversion system are available online at `http://mi.eng.cam.ac.uk/~zi201/conversions.html`.

6.1. *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. In expressing the emotions, she was asked to assume a natural, conversational style rather than a dramatic intensity. While 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.

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 [14] and smaller than the emotional prosody conversion experiments described in [15]. The numbers of

ACCEPTED MANUSCRIPT

Table 6
Number of linguistic constituents in training and test sets

	Utterances	Words	Syllables	Phones
Training Corpus	272	2170	3590	10754
Test Corpus	28	215	367	1115

words, phones and syllables in the training and test sets are given in Table 6. The mean word count per sentence is 7.9. The total duration of speech data used for training was around 15 minutes for each emotion.

6.2. *Annotations*

For each data file in the corpus and its word level transcription, a corresponding annotation file was automatically generated. Phone and state boundaries were extracted using HTK-based forced alignment [13]. The Cambridge English Pronunciation dictionary was used to identify syllable constituents for each word, as well as lexical stress positions[33]. 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 software [32] and manually corrected for mistakes.

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.

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 version 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 Figure 7.1. The sample test utterances in each emotion were changed after the first ten subjects, in order to evaluate a wider range of utterances. Table 7 displays the percentage preference rates. As can be seen, spectral conversions were consistently preferred for anger and sadness (t-test, $p << 0.01$), while for

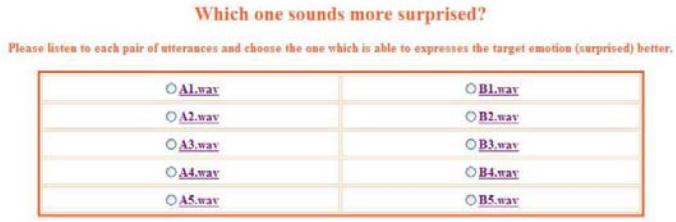


Fig. 7. The layout of the preference test

Table 7
Preference scores for GMM-based spectral conversion

	Prefer no conversion	Prefer conversion
angry	9%	91%
surprised	68%	32%
sad	13%	87%

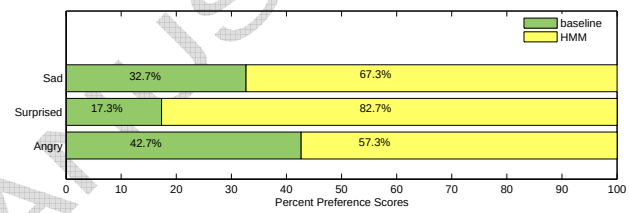


Fig. 8. Percent Preference Scores for syllable HMMs and Gaussian Normalization (baseline).

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 defined by equation 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. For both utterances, spectral conversion to the target emotion was applied. The original neutral durations were left unmodified.

30 subjects (15 male and 15 female) participated in this test. 21 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

ACCEPTED MANUSCRIPT

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 (Figure 8). Overall, the subjects strongly preferred the HMM-generated contours for surprise (t-test, $p << 0.01$). 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 ($p << 0.01$). 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.

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. 30 subjects participated in the test and each subject performed 10 comparisons per emotion. A total of 900 (30 x 10 x 3) comparisons were performed. The percentage preference scores per emotion are displayed as a stacked bar graph in Figure 9 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=0.83) 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). Overall, the shift in preferences can be explained by the fact that the stored contour segments capture more realistic F0 movements in contrast to the HMM-generated contours that are typically over-smoothed. Additionally, the incorporation of the input contour into the target cost function

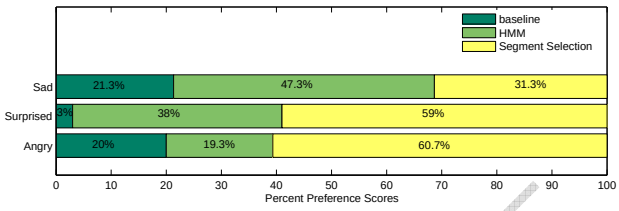


Fig. 9. Preference scores for each method and emotion

for segment selection may help select more appropriate segments (this argument is also supported by the high weight attached to this subcost for anger).

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

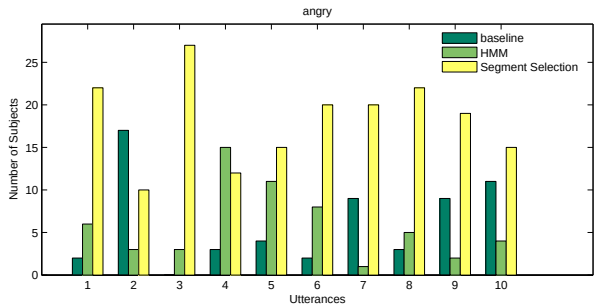
	Baseline, HMM	**Baseline, SegSel**	**HMM, SegSel**
Angry	0.83	9.4×10^{-16}	4.8×10^{-16}
Surprised	1.8×10^{-19}	1.3×10^{-28}	4.8×10^{-8}
Sad	2.3×10^{-7}	0.02	6.1×10^{-4}

The distribution of preferences across each of the ten comparisons are illustrated in Figure 10(a-c). The segment selection method was strongly preferred for all conversions to anger except utterance 2 and utterance 4. In the utterance-specific analysis of surprise (Figure 10b), 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 quite 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 (Figure 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.

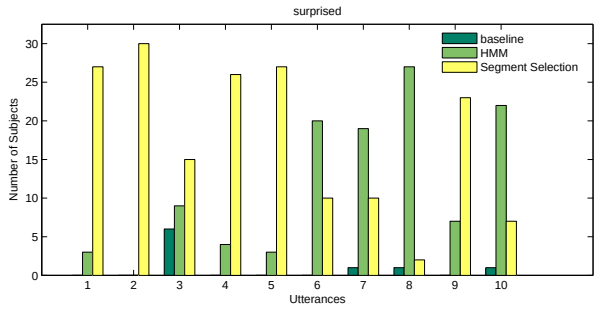
7.4. *Evaluation of Duration Conversion*

In the previous two tests, the focus was on evaluating F0 contours generated by different methods leaving the neu-

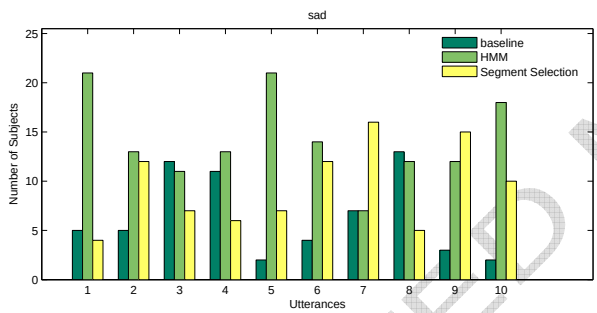
ACCEPTED MANUSCRIPT



(a) Angry



(b) Surprised



(c) Sad

Fig. 10. Utterance specific analysis of preferences across three methods of F0 conversion

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%

tral 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 Figure 7.1. 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.

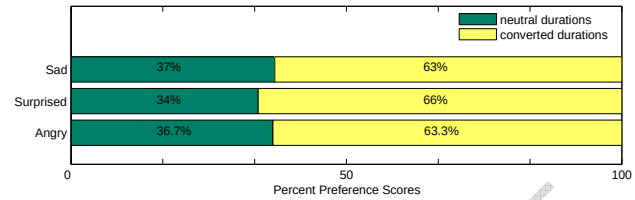
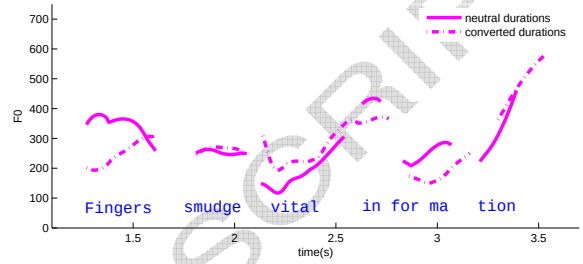
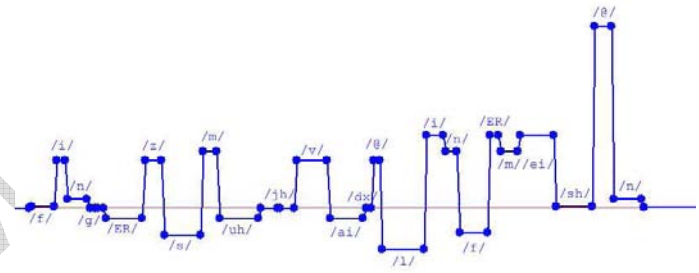


Fig. 11. Preference scores for duration conversion in each emotion



(a)



(b)

Fig. 12. Selected F0 segments for utterance "Fingers smudge vital information" before and after duration conversion (a) and the corresponding duration tier (b)

The same 30 subjects participated in the test, where each subject performed 10 comparisons per emotion (Figure 11). The results of the tests showed that converted durations were preferred more frequently than unmodified durations. This is was very significant for all emotions ($p << 0.01$). 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.

Figure 12a 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 Figure 12b, where the scaling factors for each neutral phone are identified explicitly. The horizantal line indicates a scaling factor of 1 i.e. no change. Even

ACCEPTED MANUSCRIPT

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 doubled 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 are 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, the test was first conducted using the original natural utterances of the voice talent used to record the database. Five utterances per emotion were presented to 30 subjects and the confusion matrix for this test is summarized in Table 10.

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

	Angry	**Surprised**	**Sad**	**Can't 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
Confusion scores for the emotion classification task for utterances where HMM-based contours are used

	Angry	**Surprised**	**Sad**	**Can't 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**	**Can't 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 same test was then conducted using converted neutral utterances generated by our conversion system. 10 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 (Table 11 and Table 12).

The conversion outputs using HMM-based F0 contours conveyed sadness as well as the original sad speech, while 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%). There was considerable indecision amongst subjects when classifying surprise and anger.

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 a voice-quality dominated emotion. 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

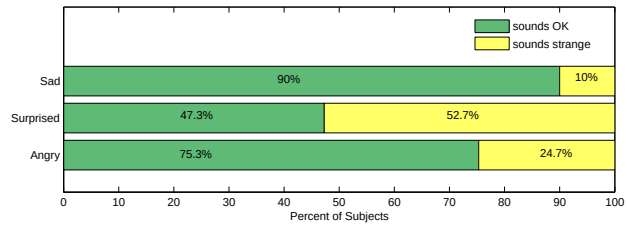


Fig. 13. *Categorical quality ratings for spectral conversion + duration conversion + HMM-based contour generation*

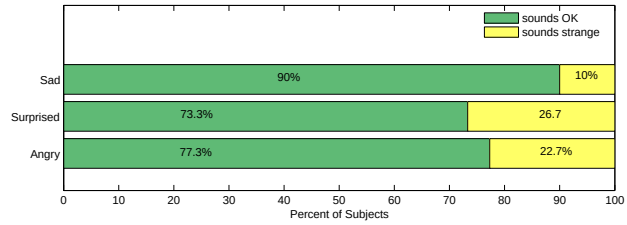


Fig. 14. *Categorical quality ratings for spectral conversion + duration conversion + F0 Segment Selection*

ACCEPTED MANUSCRIPT

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 ($p = 0.018$)

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 (Figures 13 and 14). The effect of F0 prediction method on quality was significant only in the case of surprise ($p = 0.0006$). For both methods, 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. In the surveys, a number of subjects noted that in some of the utterances, "the surprise element was there" but it was "slightly misplaced", which made them choose the "Can't decide" option. 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. 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 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 or the regression trees for duration conversion highlight the contextual factors which have the most dominant role in the expression of each target emotion. In general, surprise was found to be an emotion whose prosody is quite different from that of neutral speech and hence is highly dependent on syllable and word-level linguistic factors. On the other hand, the prosody of anger is more closely related to neutral prosody and in that case the information in the input F0 contours was a significant factor in selecting the best target contour.

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.

References

[1] Schroder, M., "Emotional Speech Synthesis - A Review", Proc. of Eurospeech, vol.1:561–564, 1999.

[2] Stylianou et al.,"Continuous Probabilistic Transform for Voice Conversion", IEEE Trans. Speech and Audio Proc. vol.6:131-142, 1998.

[3] Ye, H.,"High-Quality Voice Morphing", PhD Thesis, Cambridge University, 2005.

[4] Kain, A., Macon, M., "Spectral Voice Conversion for Text-to-Speech Synthesis", Proc. of ICASSP, vol.1:285-288, 1998.

[5] Inanoglu, Z., "Pitch Transformation in a Voice Conversion Framework", Master's Thesis, University of Cambridge, 2003.

[6] Gillett, B., King, S., "Transforming F0 Contours", Proc. of Eurospeech, 2003.

[7] Helander, E., Nurminen, J., "A Novel Method for Pitch Prediction in Voice Conversion", Proc. of ICASSP, vol.4:509-512, 2007.

[8] Yildirim, S. Bulut, M., Lee, C., Kazemzadeh,A., Busso, C., Deng, Z., Lee, S., Narayanan, S., "An acoustic Study of Emotions Expressed In Speech", Proc. of ICSLP, 2004.

[9] Barra, R., Montero, J., Arriola, J.G., Guaras, J., Ferreiros, J., Pardo, J., "On the limitations of voice conversion techniques in emotion identification", Proc. of Interspeech, 2007.

[10] Vroomen, J., Colloier, R., Mozziconacci, S., "Duration and Intonation in Emotional Speech", Proc. of Eurospeech, vol.1:577-580, 1993.

[11] Bulut, M., Lee, S., Narayanan, S. "A Statistical Approach for Modeling Prosody features using POS tags for emotional speech synthesis", Proc. of ICASSP, 2007.

[12] Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., and Shikamo, K. "GMM-based Voice Conversion Applied to Emotional Speech Synthesis", IEEE Trans. Speech and Audio Proc., 7(6):697–708, 1999.

[13] Young, S.J. *et al.* "The HTK Book Version 3.4", `http:\\htk.eng.cam.ac.uk`, Cambridge University, 2006.

[14] Wu, C.H., Hsia, C.-C., Liu, T.-E., and Wang, J.-F., "Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis", IEEE Trans. Audio, Speech and Language Proc., vol.14(4):1109–1116, 2006.

[15] Tao, J., Yongguo, K., and Li, A. "Prosody Conversion from Neutral Speech to Emotional Speech", IEEE Trans. Audio, Speech and Lang Proc., vol.14:1145–1153, 2006.

[16] Tsuzuki, H., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan, S. "Constructing emotional speech synthesizers with limited speech database", Proc. of ICSLP vol.2:1185-1188, 2004

ACCEPTED MANUSCRIPT

[17] Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T., "Modeling of various speaking styles and emotions for HMM-Based Speech Synthesis", Proc. EUROSPEECH,vol.3:2461-2464, 2003.
[18] Fallside, F., Ljolje, A., "Recognition of Isolated Prosodic Patterns using Hidden Markov Models", Speech and Language, vol.2:27-33, 1987.
[19] Jensen, U., Moore, R., Dalsgaard, P., Lindberg, B., "Modelling Intonation Contours at the Phrase Level using Continuous Density Hidden Markov Models", Computer, Speech and Language, vol.8:227-260, 1994.
[20] Tokuda, K., Zen, H., Black, A., "An HMM-Based Speech Synthesis System Applied To English.", IEEE Speech Synthesis Workshop, 2002.
[21] Ladd, D.,"Intonational Phonology", Cambridge University Press, 1996.
[22] Banziger, T., Scherer, K., "The Role of Intonation In Emotional Expressions", Speech Communication, vol.46:252-267, 2005.
[23] Inanoglu, Z., Young, S., "Intonation Modelling and Adaptation for Emotional Prosody Generation", Proc. of Affective Computing and Intelligent Interaction, 2005.
[24] Ross, K., Ostendorf, M., "A dynamical system model for generating F0 for Synthesis", Proc. of ESCA/IEEE Workshop on Speech Synthesis, 1994.
[25] van Santen, J.P.H, Hirschberg, J., "Segmental effects on timing and height of pitch contours", Proc. of ICSLP, 1994.
[26] Tokuda, K., Masuko, T., Miyazaki, N., Kobayashi, T., "Multi-Space Probability Distribution HMM", IEICE Trans. Inf. and Systems,vol.3:455-463, 2002.
[27] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T, Kitamura, T., "Speech Parameter Generation Algorithms for HMM-Based Speech Synthesis", Proc. of ICASSP,vol.3:1315-1318, 2000.
[28] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T, Kitamura, T., "Duration Modelling for HMM-Based Speech Synthesis", Proc. of ICSLP, 1998.
[29] Toda, T., Tokuda, K., "Speech Parameter Generation Algorithm Considering Global Variance for HMM-Based Speech Synthesis", Proc. of Interspeech, 2005.
[30] Tian, J., Nurminen, J., Kiss, I., "Novel Eigenpitch-based Prosody Model for Text-to-Speech Synthesis", Proc. of Interspeech, 2007.
[31] Inanoglu, Z., "Data-driven Parameter Generation for Emotional Speech Synthesis", PhD Thesis, University of Cambridge, 2008.
[32] http://www.fon.hum.uva.nl/praat/
[33] Jones, D., "Cambridge English Pronunciation Dictionary", Cambridge University Press, 1996.
[34] Inanoglu, Z., Young, S., "A System for Transforming the Emotion in Speech: Combining Data-Driven Conversion Techniques for Prosody and Voice Quality", Proc. of Interspeech, 2007.
[35] Inanoglu, Z., "Data-driven Parameter Generation for Emotional Speech Synthesis", PhD Thesis, University of Cambridge, 2008.

ACCEPTED MANUSCRIPT