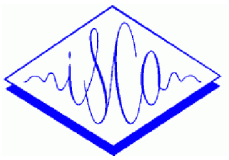
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Exploring Rich Expressive Information from Audiobook Data Using Cluster Adaptive Training

Langzhou Chen, Mark J.F. Gales, Vincent Wan, Javier Latorre, Masami Akamine

Toshiba Research Europe Ltd., Cambridge, UK
lchen,mjfg,vincent.wan,javier.latorre@crl.toshiba.co.uk, masa.akamine@toshiba.co.jp

Abstract

Audiobook data is a freely available source of rich expressive speech data. To accurately generate speech of this form, expressiveness must be incorporated into the synthesis system. This paper investigates two parts of this process: the representation of expressive information in a statistical parametric speech synthesis system; and whether discrete expressive state labels can sufficiently represent the full diversity of expressive speech. Initially a discrete form of expressive information was used. A new form of expressive representation, where each condition maps to a point in an expressive speech space, is described. This cluster adaptively trained (CAT) system is compared to incorporating information in the decision tree construction and a transform based system using CMLLR and CSMAPLR. Experimental results indicate that the CAT system outperformed the contrast systems in both expressiveness and voice quality. The CAT-style representation yields a continuous expressive speech space. Thus, it is possible to treat utterance-level expressiveness as a point in this continuous space, rather than as one of a set of discrete states. This continuous-space representation outperformed discrete clusters, indicating limitations of discrete labels for expressiveness in audiobook data.

Index Terms: expressive speech synthesis, hidden Markov model, cluster adaptive training, audiobook

1. Introduction

Expressive speech synthesis is a challenging topic in current text to speech synthesis (TTS) research. This paper aims to improve expressive TTS systems for audiobook data. This data is very rich in expressive information as the readers aim to produce lively, animated, stories. How to make use of the audiobook data for training TTS system was described in [1] for example. However it is non-trivial to obtain expressive information for this data, as manual annotation is expensive and yields poor inter-annotator agreement. To address this issue unsupervised clustering approaches for audiobook data have been proposed [2, 3]. This makes no assumptions about the exact form of expressive labels, just that expressive information can be represented as one of a set of discrete labels.

This expressive information must then be introduced into the acoustic models of the parametric statistical synthesis system. Since there is typically insufficient data to robustly training individual expressive state models, decision-tree [4] and transform-based approaches [5] have been investigated for both emotion labelled data [4] and for unsupervised audiobook expressive states [3], average expressive state system (AESS).

In this work an alternative approach based on representing the expressive state as a point in a multi-dimensional continuous space is investigated. Here the point in space for each of the expressive states and the definition of the multi-dimensional space are based on the cluster adaptive training (CAT) approach, that has previously been used for polyglot TTS [6]. This form extends the original CAT formulation [7] to allow separate decision trees to be constructed for each of the clusters. This yields a more complex expressive speech space to be defined as any changes in the context-dependency of the speech with expressiveness can be modelled. Any dependence of the decision tree on the style cannot be modelled in AVM-style systems [5] as these only allow a single decision tree to be used.

Another interesting attribute of CAT systems is that the quantity of data required to estimate a point in the expressive speech space is very small. Though this is not important if expressiveness is just represented as one of a discrete set of expressive labels, if a more detailed expressive representation is required it may be useful. This paper also investigates whether a richer expressive representation based on representing the expressive state for individual speech utterances as points in the multi-dimensional expressive space is useful. This will be referred to as utterance-based CAT. Although CAT requires little data to determine the expressive point in space, data sparseness may still be a problem for short utterances, especially for expressive points associated with duration. This paper describes a method to use expressive labels as prior information to smooth the utterance-based estimators [8]. This allows a balance between expressiveness and robustness to be achieved in CAT-style representations.

The purpose of this work is to investigate representations of expressiveness, whether discrete or continuous, and the incorporation of this information into a TTS system for audiobook data. To avoid issues with mapping from text to expressive state, this work assumes that the expressive state can be reliably obtained. Thus here audio data is used to obtain the expressiveness in the same fashion as [3]. Approaches for implementing the mapping from text to expressive state will be investigated in future work.

2. Expressive CAT Model

Originally, CAT was developed for speech recognition to enable rapid speaker adaptation [7]. Here each speaker is associated with a point in a speaker space. Both the specification of the space and estimation of the points for each can be obtained using maximum likelihood (ML). CAT has been extended for statistical parametric synthesis to perform the speaker and language factorization [6]. There the language of the speaker is represented as a point in a language-space. Since statistical synthesis approaches incorporate significant information into the decision tree, the TTS version of CAT allows separate decision trees to be specified for each cluster. This is the form of CAT used in this work. To apply CAT to expressive speech synthesis,

the expressive state associated with an audiobook utterance is represented as a point in a multi-dimensional expressive-speech space. Note as TTS systems make use of multiple streams, each stream will have its own multi-dimensional space and associated point in that space.

The CAT model consists of a set of cluster models. Each cluster model contains a set of Gaussian mean parameters while the Gaussian variances are shared over all clusters. When this CAT model is used to calculate the likelihood of an observation vector, the mean vector to be used is a linear interpolation of all the cluster means, i.e.

$$p(\boldsymbol{o}_t|\boldsymbol{\lambda}^{(e)}, \mathbf{M}^{(m)}, \boldsymbol{\Sigma}^{(m)}) = \mathcal{N}(\boldsymbol{o}_t; \mathbf{M}^{(m)}\boldsymbol{\lambda}^{(e)}, \boldsymbol{\Sigma}^{(m)}) \quad (1)$$

where $\mathbf{M}^{(m)}$ is the matrix of P cluster mean vectors for component m, i.e. $\mathbf{M}^{(m)} = \begin{bmatrix}\boldsymbol{\mu}^{(m,1)} & \ldots & \boldsymbol{\mu}^{(m,P)}\end{bmatrix}$ and $\boldsymbol{\lambda}^{(e)}$ is the point in expressive state. It is simple to extend this form of representation to include multiple regression classes with each of the expressive states. In common with standard CAT approaches the first cluster is specified as a bias cluster, thus

$$\boldsymbol{\lambda}^{(e)} = \begin{bmatrix}1 & \lambda_2^{(e)} & \ldots & \lambda_P^{(e)}\end{bmatrix}^{\mathsf{T}} \quad (2)$$

In this work, the CAT model is used to model the expressiveness of audiobook data. The bias cluster model represents the expression independent factors of the training data, while the non-bias clusters will be used represent the expression dependent factors of the training data.

The training process of the CAT model can be divided into three main parts: decision tree construction; multi-dimensional expressive state definition; and the estimation of the expressive-state in the multi-dimensional space. For this work a simple initialisation scheme was used based on an expressive-independent system. This system was used to initialise the bias cluster, with all other clusters being set to zero means. This allows the decision tree construction and parameter estimation in [6] to be used. Due to lack of space it is not possible to describe in detail all the stages of the the CAT build. This section will concentrate on the estimation of the CAT weights as this will be used as the expressive speech representation.

The point in the multi-dimensional expressive space will be given by the CAT weight. The ML estimate of this weight $\hat{\boldsymbol{\lambda}}$, ignoring the regression class, can be found using the following auxiliary function

$$\mathcal{Q}(\hat{\boldsymbol{\Lambda}}; \boldsymbol{\Lambda}) = \hat{\boldsymbol{\lambda}}\mathbf{k} - \frac{1}{2}\hat{\boldsymbol{\lambda}}^{\mathsf{T}}\mathbf{G}\hat{\boldsymbol{\lambda}} + D \quad (3)$$

where D represents all the terms independent to $\hat{\boldsymbol{\lambda}}$. The sufficient statistics $\mathbf{G}$ and $\mathbf{k}$ are given by

$$\mathbf{G} = \sum_{m,t} \gamma_t^{(m)} \mathbf{M}^{(m)\mathsf{T}} \boldsymbol{\Sigma}^{(m)\text{-}1} \mathbf{M}^{(m)} \quad (4)$$

$$\mathbf{k} = \sum_{m} \mathbf{M}^{(m)\mathsf{T}} \boldsymbol{\Sigma}^{(m)\text{-}1} \sum_{t} \gamma_t^{(m)} (\boldsymbol{o}_t - \boldsymbol{\mu}^{(m,1)}) \quad (5)$$

where $\gamma_t^{(m)}$ is the occupancy probability of component m at time t, $\boldsymbol{\mu}^{(m,1)}$ is the mean vector of component m from bias cluster. In (4) and (5), $\mathbf{M}^{(m)}$ is the matrix of non-bias cluster mean vectors for component m, and $\hat{\boldsymbol{\lambda}}$ to be calculated is the non-bias cluster weight, since the weight of bias cluster is fixed to 1.

Differentiating the auxiliary function w.r.t $\hat{\boldsymbol{\lambda}}$ and equating to zero yields,

$$\hat{\boldsymbol{\lambda}} = \mathbf{G}^{\text{-}1}\mathbf{k} \quad (6)$$

In this CAT model for expressive TTS systems, the expressiveness information is recorded in the CAT weight vector, which corresponds to a point in the multi-dimensional expressive space. There is a choice in how this point in space is defined. In [3] discrete labels, derived in an unsupervised fashion, were used. The same approach can be adopted for CAT-style systems. In this case the summation over t in (4) and (5) is for all frames that are allocated to a particular discrete label.

However, the expressiveness in audiobook data is very rich. Using a set of discrete labels may not be sufficient to cover the full diversity of audiobook data. A more accurate representation for the expressiveness of each utterance may be beneficial. An alternative approach is to represent each utterance as an individual point in the continuous CAT expressive space. In this case the summation over t in (4) and (5) is only for the frames of a an utterance.

Although CAT weight estimation does not require large amounts of adaptation data, for utterance-based CAT, robustness of the estimates may be a problem for the duration weight. For the durations the observation unit is a state rather than a frame. Introducing prior information is a standard approach to handle such problems. It has been used successfully in transformation estimation for rapid speaker adaptation in the ASR domain [8]. In the TTS domain, [9] proposed to use prior information to smooth CMLLR transformations in a tree structure. In this work the count smoothing approach described in [8] is used. Here the sufficient statistics $\mathbf{G}$ and $\mathbf{k}$ in section 2 are computed at the label level. This is assumed to yield robust estimates. These statistics are then smoothed with the utterance level counts. Thus

$$\mathbf{G} = \mathbf{G}_{\mathtt{utt}} + \tau \frac{\mathbf{G}_{\mathtt{lab}}}{\sum_{m,t} \gamma_t^{(m)}}, \quad (7)$$

$$\mathbf{k} = \mathbf{k}_{\mathtt{utt}} + \tau \frac{\mathbf{k}_{\mathtt{lab}}}{\sum_{m,t} \gamma_t^{(m)}} \quad (8)$$

where $\mathbf{G}_{\mathtt{utt}}$, $\mathbf{k}_{\mathtt{utt}}$ are the utterance level statistics and $\mathbf{G}_{\mathtt{lab}}$, $\mathbf{k}_{\mathtt{lab}}$ are the label-level statistics. The summation over t is for all the frames associated with the expressive label. τ is the weight which controls the contribution of prior statistics to be τ frames. The weight vector can then be computed using (6).

By appropriately weighting this smoothing a balance can be achieved between expressiveness and robustness. As the main aim of the smoothing is to address the duration estimation, the weight of prior statistics τ was set to a relatively small value, 5. This value was not tuned for the particular task.

3. Experimental Results

Preliminary experiments were based on the audiobook data "A Tramp Abroad", read by John Greenman. This book contains 56 chapters, which were divided into 51 chapters for training and 5 for evaluation. The data was processed using lightly supervised techniques [1]. The data was segmented into 3 types of speech units, or utterances: narration, carrier and direct speech [3]. This yielded 4.8k utterances with a 100% word accuracy against the book text for model training. The average length of a training utterance was 6.8 seconds. The sampling rate of the data was 16kHz and acoustic features consisted of 40 mel-cepstral

coefficients, logF0, 21 (approximately bark scaled) BAP plus their delta and delta-delta information. The models were 5 state left-to-right multi-space probability distribution hidden semi-Markov models.

The CAT model used in this work comprised five clusters; one bias cluster and four non-bias clusters. Discrete expressive labels were obtained by using an unsupervised hierarchical k-means clustering approach. The data and expressive labels were the same as those used in [3]. Initially a CAT system based on these discrete automatically derived expressive labels was trained. Additionally, to investigate the use of a continuous multi-space representation, a CAT system using a continuous expressive representation for each utterance was also built.

The CAT training process for discrete labels is summarised below.

1. Construct an expression independent system, set this to be the bias cluster.
2. Using the hierarchical k-means trees allocate each discrete label to one of 4 clusters.
3. For each cluster set $\boldsymbol{\lambda}$ to one for that cluster and the current bias cluster and zero otherwise.
4. Construct the decision tree for each cluster given the current model parameters.
5. For each of the discrete labels estimate the point in the current multi-dimensional CAT expressive space (weight vector).
6. Update cluster models given the current trees and discrete label positions.
7. If parameter and weight estimation not converged goto step 5.
8. If tree estimation not converged goto step 4.

In addition, it is possible to optimise the CAT system using an expressive representation based on a speech unit level point in a multi-dimensional continuous space. Here the process above can be repeated. However step 5 is replaced by computing a point for each of the speech units, rather than each label. Again it is possible to iterate updating the model parameters, points in space (weights) and decision trees. In this work, the decision tree update was not performed for this system. Thus the utterance-based CAT model shared the decision trees with the label-based CAT model, but with different cluster parameters.

Initially the label-based expressive representation with CAT was compared with two contrast systems, initially presented in [3]. The first uses decision trees, where the expressive labels are used as questions in the decision tree construction, labelled DT. The other is an average expression model based on CMLLR/CSMAPLR transforms, the AESS system. Note for both these approaches it is only straightforward to classify expressiveness as one of a set of discrete labels.

The synthesis evaluations were based on 75 test utterances from 5 test chapters of audiobook, including 40 narration utterances, 25 direct speech utterances and 10 carrier utterances. The listening tests were crowd-sourced via CrowdFlower. In common with the experiments in [3] the cluster labels during synthesis were given, rather than having to be derived from the text. Two aspects of the synthesis performance were investigated; expressiveness and voice quality. ABX tests were used to evaluate whether the expressiveness was well represented, and preference tests were used to evaluate the voice quality.

Table 1: ABX test results for expressiveness

DT	AESS	CAT	p
44.4%	55.6%		0.001
	39.9%	60.1%	$<$0.001
41.1%		58.9%	$<$0.001

Table 2: Preference test results for voice quality

DT	AESS	CAT	no preference	p
43.2%	48.3%		8.5%	0.100
32.4%		54.7%	12.9%	$<$0.001
	35.5%	50.9%	13.6%	$<$0.001

The expressiveness test results and voice quality test results are given in table 1 and table 2 respectively. These indicate that the CAT based system outperformed the decision tree method and AESS method significantly in both expressiveness and voice quality.

Though the CAT-based system yielded gains over the two alternative approaches, it was not clear to what extent the use of discrete labels limited the expressiveness of the system. To evaluate this the continuous-space expressive representation was used. Again the point in expressive-space was assumed to be known. In this work this was obtained by projecting the acoustics associated with a test utterance to a point in the expressive space. Note, though this mapping is more detailed, and would be harder than mapping text to an expressive label, the aim of these experiments is to examine possible advantages of continuous space representations.

Initially attributes of this continuous expressive-space were investigated. Using a continuous expressive space, each training utterance is represented as a point. Ideally, if the distance between two utterances is close in CAT weight space, they should contain similar expressiveness. This should then be reflected in the perception of the utterance.

The first experiment was to evaluate this consistency between distance in CAT weight space and human perception. In [3], four different sets of acoustic features, labelled A, B, C and D, were compared for the unsupervised clustering of the audiobook data. Each feature led to a different classification of the training data. To evaluate the performance of different classifications, a listening test was carried out. The listeners listened to a reference utterance from one of two clusters, and were then asked to say which of two test utterances (one from the reference, one from another cluster) had a similar expressiveness.

Table 3: Subjective and objective scores for expressive clustering

feature set	A	B	C	D
% correct	71	75	86	70
J-lf0	-9.10	-8.98	-5.79	-8.59
J-mcep	-9.57	-9.96	-9.79	-11.13
J-bap	-11.84	-11.80	-10.17	-12.59

The results for the listening tests from [3] are given in the row labelled *% correct* in table 3. This indicates that clustering based on feature C yields clusters that are more distinguishable

in terms of subjective perception.

The *% correct* measure can be viewed as a subjective between-to-within variance measure i.e. the *% correct* results in table 3 describe two aspects of classification process, whether data with the same label exhibit similar expressive characteristics and whether data in different labels are distinguishable. Rather than requiring subjective measures to determine these attributes of the labels, it would be preferable to use objective criteria. One such criterion that can be used with the utterance level points in the CAT expressive space is related to linear discriminant analysis

$$J = \log(|\mathbf{S}_{\mathtt{w}}^{-1}\mathbf{S}_{\mathtt{B}}|) \quad (9)$$

where $\mathbf{S}_{\mathtt{B}}$ and $\mathbf{S}_{\mathtt{w}}$ are the average between class scatter matrix and within class scatter matrix for the weight vectors of the training utterances respectively. The value of J for three sets of parameters are given in table 3. LogF0 (*J-lf0*) is expected to contain significant expressive information. This is clearly indicated by the high value associated with the feature C labels. In addition results for Mel-Cep (*J-mcep*) and BAP (*J-bap*) are shown. Though the benefit of using feature C clusters is not as clear as for logF0, the overall trend is that feature C clusters are better than the other features. This is consistent with the subjective results. This indicates that the continuous expressive spaces defined by CAT may also be useful for comparing, and clustering, expressive representations for speech synthesis.

Table 4: ABX results for utterance-based CAT

AESS	CAT		p
	label	utterance	
39.9%	60.1%		$<$0.001
34.0%		66.0%	$<$0.001
	38.7%	61.3%	$<$0.001

Since the expressiveness in audiobook data is highly variable, a continuous-space representation may be useful in the synthesis process. Table 4 gives the expressiveness performance of the continuous utterance-based CAT system against the AESS and the original discrete expressive label CAT system. The results indicate that the continuous space does yield significant gains over the contrast systems. Note for the utterance based system, priors, based on the labels, were used when determining the point in expressive-space. Without this prior information utterance-based CAT was still better than label-based CAT in an ABX test. However, the difference was not significant. The performance gains of the continuous expressive representation, compared to the discrete label based approach, indicates that the expressiveness of audiobook data may be too varied to be fully modelled with discrete expressive labels. Table 4 also gives the comparison results between the utterance-based CAT system and the AESS system. Not surprisingly, the utterance-based CAT system outperformed the AESS system significantly. The ability to model the detailed expressiveness of speech data as a point in a multi-dimensional continuous space is one advantage of the CAT approach. It is not practical for methods like AESS because of the sparseness of the training data and the high computation cost.

It is possible to also assess the quality of utterance-based CAT versus the discrete labelled based scheme. However in preliminary experiments the results were highly varied. Utterance based CAT tended to produce far more varied, and expressive speech. Without context information, listeners tended to select more neutral sounding speech. Highly expressive speech, without appropriate context, will tend to have lower quality than neutral speech. Appropriate ways to assess voice quality in expressive synthesis will be examined in future work.

4. Conclusion

This work trained an expressive TTS system using audiobook data, based on cluster adaptive training (CAT) technology. Similar to the average expression speech synthesis (AESS) method, CAT uses a transformation to represent expressiveness. This allows both approaches to avoid the data fragmentation that occurs when using expressive specific models, or expressive decision tree questions. CAT uses a simple representation for expressiveness, a point in a multi-dimensional continuous space. However by allowing cluster-specific decision trees this space offers a rich choice of possible synthesis model parameters. Preliminary experiments were based on known expressive labels. For discrete expressive labels, derived by unsupervised clustering, CAT-based models significantly outperform decision tree and the AESS methods in both expressiveness and voice quality tests.

To model the full variety of expressiveness in audiobook data, discrete labels may not be sufficient. CAT-style models allow a continuous expressive representation. In this work each utterance was mapped to a point in the multi-dimensional expressive space. Experiments show that this richer continuous representation yields significantly more expressive systems than the discrete labels.

The current experiments have assumed that the labels, or point in expressive space, are known. Future work will concentrate on deriving this information automatically from text.

5. References

[1] N. Braunschweiler, M.J.F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. of Interspeech*, 2010, pp. 2222–2225.

[2] E. Szekely, J. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering expressive speech styles in audiobooks using glottal source parameters," in *Proc. of Interspeech*, 2011.

[3] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M.J.F. Gales, and K. Knill, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. of ICASSP*, 2012.

[4] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. on information and systems*, vol. E88-D, pp. 503–509, 2005.

[5] J. Yamagishi, T.Kobayashi, M.Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. of ICASSP*, 2007.

[6] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Transactions on Audio Speech and Language Processing*, to appear.

[7] M. J. F. Gales, "Cluster adaptive training of hidden markov models," *IEEE Transactions on Audio Speech and Language Processing*, vol. 8, pp. 417–428, 2000.

[8] C. Breslin, K. Chin, M.J.F. Gales, K. Knill, and H. Xu, "Prior information for rapid speaker adaptation," in *Proc. of Interspeech*, 2010.

[9] J. Yamagishi, T.Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation method," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 66–83, 2009.