UNSUPERVISED CLUSTERING OF EMOTION AND VOICE STYLES FOR EXPRESSIVE TTS

Florian Eyben*, Sabine Buchholz, Norbert Braunschweiler,
Javier Latorre, Vincent Wan, Mark J. F. Gales, Kate Knill

Toshiba Research Europe Ltd., Cambridge Research Lab, 208, Cambridge Science Park,
Cambridge, CB4 0GZ, UK

eyben@tum.de, sabine@sbuchholz.eu, {norbert.braunschweiler@crl.toshiba.co.uk,
javier.latorre,vincent.wan,mjfg,kate.knill}@crl.toshiba.co.uk

ABSTRACT

Current text-to-speech synthesis (TTS) systems are often perceived as lacking expressiveness, limiting the ability to fully convey information. This paper describes initial investigations into improving expressiveness for statistical speech synthesis systems. Rather than using hand-crafted definitions of expressive classes, an unsupervised clustering approach is described which is scalable to large quantities of training data. To incorporate this "expression cluster" information into an HMM-TTS system two approaches are described: cluster questions in the decision tree construction; and average expression speech synthesis (AESS) using cluster-based linear transform adaptation. The performance of the approaches was evaluated on audiobook data in which the reader exhibits a wide range of expressiveness. A subjective listening test showed that synthesising with AESS results in speech that better reflects the expressiveness of human speech than a baseline expression-independent system.

Index Terms— Expressive synthesis, text-to-speech, unsupervised clustering, Average Voice Model, HMM-TTS

1. INTRODUCTION

Proper use of expression makes speech more interesting and aids understanding of content by adding nuances and information beyond the pure text content. Expressiveness in speech includes emotions (e.g. angry, sad), speaking style (e.g. whisper, boasting), and the "character voices" often used in story reading (e.g. "old man"). While state-of-the-art TTS systems achieve high intelligibility and naturalness for reading isolated sentences in a relatively neutral style, the synthesis of expressive speech is still a challenge.

The first key components in expressive text-to-speech synthesis (ETTS) are the speech corpus used and the annotation of "expression" in that corpus. In most approaches to date a separate speech corpus was recorded for each expression of interest [1]. However, creating such dedicated speech corpora is costly and time-consuming, and typically limited to a handful of expressions, lacking generalisation. Researchers have started to explore using audiobooks e.g. [2, 3, 4] as a source of training data for TTS due to the wide variety of expressions contained. This paper presents an approach to derive ETTS from audiobook data.

Unlike dedicated emotional corpora, the emotions and speaking styles are mixed in audiobooks. There is no standard set of classes for annotating such data so manual annotation is highly subjective, with poor inter-annotator agreement. The size of such corpora (10-20 hours) also makes annotation of multiple books impractical in terms of time and cost. Therefore an unsupervised clustering approach is taken here to produce automatic expression annotations. The aim of the clustering is to place similar (emotion and style) short book units into the same or proximate clusters. Recently Szekely et al. [5] have proposed self-organising feature maps to cluster audiobook data based on voice quality (but no TTS was built). For this work, hierarchical k-means clustering is used based on acoustic features derived from work in emotion recognition [6]. This approach was chosen due to its simplicity and linear scalability.

Once the expressive clusters are learnt, ETTS can be trained. HMM-based speech synthesis is used to provide a flexible framework to model the varying expressions. Two approaches are investigated: incorporating the cluster labels as context features within the decision tree creation process cf. [7]; and average voice speech synthesis [8] with the expressive clusters acting as "speakers".

The final step in realising ETTS is to determine the appropriate expression at synthesis. There are 3 main scenarios for expression derivation: from text (audiobook reading); given by external sources (in a dialog system or manually specified); and from audio (speech-to-speech translation). The unsupervised clusters do not have a human-readable label such as "sad". However, a machine learner could be trained to predict the correct expressive cluster ([9]). Alternatively, for the manual specification case a user might get a "feel" for the clusters by listening to synthesised examples, possibly assigning their own, informal labels, prior to selecting the cluster for a specific utterance. In this paper the issue of how to choose the appropriate cluster is not investigated. It is assumed that the cluster can be chosen reliably.

The next sections are as follows: Section 2 describes the proposed clustering method for unsupervised labelling of the expressions. The expressive speech synthesis approaches are described in Section 3. Experimental results are presented in Section 4. Finally conclusions are drawn in Section 5.

2. UNSUPERVISED CLUSTERING OF EXPRESSIONS

Since there is no agreement of what the optimum set of expressive annotation labels for an audiobook is, unsupervised clustering is used to place similar (emotion, speaking style and character voices) book units into the same or proximate clusters. To cluster the training data into expressions three questions need to be addressed: (i) the linguistic level at which the features will be labelled, (ii) the nature of the features, and (iii) the clustering approach.

Emotion recognition and classification is typically performed at

*Florian Eyben performed this work while an intern at Toshiba CRL. He is a PhD student at Technische Universität München, Germany.

the sentence level. This is because the longer the unit of analysis is, the more stable the result provided the emotion does not change [10]. A lot of the expressive speech in audiobooks occurs within direct speech whereas the associated carrier phrases, such as 'he said', tend to be neutral in style. Therefore for clustering, the audiobook sentences were sub-divided into three types of units: narration, carrier and direct speech. The presence of double quotation marks in the audiobook text was used to detect direct speech [1]. For example, a sequence such as:

He said angrily, "Yes! I know. Bye!",
and hung up. Then he left.

would be split into the 6 underlined units: an initial *carrier* unit (a part not in quotes of a sentence containing quotes), three *direct speech* units, another *carrier*, and a *narration* unit (a complete sentence not inside or containing quotes). In this example, the reader would be expected to convey the angry tone only in the direct speech units.

There are three primary options in terms of the feature sets that can be used for the unsupervised clustering: acoustic-only; text-only; acoustics and text. In synthesis generally only the text will be available. Typically for emotion recognition and classification only acoustic features are used, although there has been some very recent work on combining with text features [9]. For simplicity only acoustic features were considered in this work. The feature extraction followed the commonly used supra-segmental modelling approach. In this a single, high-dimensional feature vector is extracted for a speech unit of variable length (e.g. a sentence) e.g. [11]. Low-level acoustic features are extracted on a frame level and mapped to the unit level via functionals (mean, standard deviation, etc.).

To derive the feature vector, the Interspeech 2011 Speaker State Challenge set [6] was used as a starting point. This has 4,368 features. The set includes a number of items which are not (or poorly) related to expressiveness, such as spectral features. Therefore, the list of low-level descriptors was reduced to the following prosodic descriptors: F_0, voicing probability, loudness, voice quality (local jitter and shimmer, logarithmic HNR). For these descriptors, functionals including arithmetic mean, flatness, standard deviation, and skewness were applied to yield a set of 163 features, called *featA*.

Initial experiments using *featA* indicated a link between utterance length and cluster assignment. This was found to be due to some features in *featA* being highly correlated with the utterance length. A new set of features was produced by removing all features with a correlation coefficient (wrt. utterance length) higher than 0.2. A total of 69 features remained in this set, *featB*. Motivated by [5], two manually selected feature sets were also created with a minimal set of prosodic and voice quality descriptors and their functionals:

featC 8 features: mean of F_0, voicing probability (p_v), local jitter and shimmer, and logarithmic HNR; standard deviation of F_0; mean of absolute delta of F_0 and p_v.

featD 4 features: mean of F_0 and p_v; mean of absolute delta of p_v; standard deviation of F_0.

Before clustering all feature vectors were standardised to have zero mean and unit variance. This ensures that all features are equally accounted for and that badly scaled features do not bias the clustering.

An unsupervised clustering approach is required. For this work hierarchical k-means clustering, similar to the x-means algorithm described in [12], was applied in a cascade of hierarchical binary splits. The number of leaf clusters was controlled using BIC and a minimum cluster occupancy criterion of 20. A maximum tree depth of 5 levels limited the maximum number of leaf clusters to 32. A Euclidean distance metric was used to determine which clusters to split. To improve the stability of the algorithm, the initial cluster centres in each split were initialised heuristically with a small perturbation to the left and right of the original centroid.

[1] The majority of quoted material is direct speech. However, there are also "other" uses of quotes. In the present work, these uses were not distinguished.

3. EXPRESSIVE SPEECH SYNTHESIS

Once the expressive clusters are generated, a speech synthesis system may be trained. HMM-based speech synthesis is used to provide a flexible framework to model the varying expressions. Each unit of the acoustic data is assigned to a leaf node of the hierarchical expressive cluster tree. One approach would be to train individual models for each cluster leaf node. However, some clusters have very little data associated with them so the models produced would be poor. An alternative is to incorporate the expressive cluster assignments as context features [7] allowing questions to be asked about them in the decision tree generation process. This allows a balance of quantity of data and context sensitivity to expressiveness. The decision tree question set is extended to include questions about the clusters and a mixed-expression model trained using an otherwise standard mono-speaker training process. Questions are asked about all nodes in the hierarchical cluster tree, allowing broader expressive classes to be incorporated into the decision trees. In the $\log F_0$ trees of the *decision tree* system in Section 4, the path of 98% of the full-context phones includes these questions for at least one state.

A problem with using the clusters as questions in the decision trees is that it still fragments the training data. To prevent this an adaptation based approach can be applied. There are a number of schemes that could be used. For this work the CMLLR/CSMAPLR average voice speech synthesis [8] approach is adopted to perform average expression speech synthesis (AESS). For AESS no cluster questions are used in the decision tree generation. An expression-independent, full context maximum likelihood model is trained in the standard way. Speaker adaptive training based on CMLLR is then applied with each expression cluster acting as a "speaker". Here, the likelihood of the observation $\boldsymbol{o}(t)$, which is uttered using expression e, from state i can be expressed as

$$p(\boldsymbol{o}(t)|i,e,\mathcal{M}) = \left|\boldsymbol{A}^{(e)}_{r(i)}\right| \mathcal{N}\left(\boldsymbol{A}^{(e)}_{r(i)}\boldsymbol{o}(t) + \boldsymbol{b}^{(e)}_{r(i)} \;;\; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right), \quad (1)$$

where $e \in \{1,\ldots,E\}$ are expression indices, E corresponds to the total numbers of expression clusters, and $r(i) \in \{1,\ldots,R\}$ is the regression class for state i. During training the set of state parameters, $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$, and the set of CMLLR expression transforms, $\{\boldsymbol{A}^{(e)}_r, \boldsymbol{b}^{(e)}_r\}$ are estimated. Block diagonal global CMLLR transforms are initially used for speech, silence and pause and the decision trees regenerated. Regression class CMLLR transforms are then trained with the decision trees held fixed and the model parameters updated. The state-duration distributions are treated in the same fashion.

To obtain the models and transforms for synthesis, each expression cluster is treated as a "target speaker". The training data for each cluster is reused as the adaptation data. The CMLLR expression transforms from training are refined using CSMAPLR to yield $\{\hat{\boldsymbol{A}}^{(e)}_r, \hat{\boldsymbol{b}}^{(e)}_r\}$. This is followed by MAP adaptation of the means. The final MAP adapted means are given by:

$$\hat{\boldsymbol{\mu}}^{(e)}_i = \frac{\tau\boldsymbol{\mu}_i + \sum_{t\in t(e)}\gamma_t(i)\left(\hat{\boldsymbol{A}}^{(e)}_{r(i)}\boldsymbol{o}(t) + \hat{\boldsymbol{b}}^{(e)}_{r(i)}\right)}{\tau + \sum_{t\in t(e)}\gamma_t(i)} \quad (2)$$

where $\boldsymbol{\mu}_i$ is the mean vector of the state output distribution of the average expression model and $\gamma_t(i)$ the state occupancy probability, $t(e)$ is all the data associated with expression e, and τ is the hyper-parameter [2]. At synthesis this MAP adapted model is combined with the CSMAPLR expression transforms.

4. EXPERIMENTAL RESULTS

4.1. Database

A publicly available audiobook was used for these experiments. The audiobook is based on the book "A Tramp Abroad" written by Mark Twain and read by John Greenman [3]. This audiobook was chosen as it includes a large variety of emotions, speaking styles and character voices. The book contains 56 chapters and total running time is 15 hours and 46 minutes. The reader uses different speaking styles for various characters in the book, e.g. "raven speech", "German lady", and he also uses a wide range of emotions and speaking styles to express funny, bizarre, etc. passages of the book. 7,055 sentences were extracted from the audiobook (for which the orthographic book sentence and corresponding audio could be aligned) using the segmentation part of the lightly supervised sentence segmentation and selection approach [4]. These sentences were further split, using double quotation marks as delimiters, into 7,371 units (616 carrier, 1,725 direct speech, 5,030 narration).

4.2. Unsupervised clustering and its evaluation

All 7,371 units were clustered using the method outlined in Section 2 for each of the four feature sets described. The number of leaf clusters created and the minimum and maximum number of units per leaf cluster can be seen in Table 1. As expected, larger feature sets result in more clusters in total, but fewer extremely large ones.

Feature set	# features	# clusters	# units /cluster	% correct cluster
featA	163	28	49–745	71
featB	69	27	24–713	75
featC	8	20	43–1136	86
featD	4	13	29–1505	70

Table 1. Results from clustering and evaluating four feature sets: number of resulting clusters, range of number of units per cluster, percentage of listening test answers where listeners chose the file from the same cluster as the reference file.

To investigate the influence of feature set choice on clustering performance, a subjective evaluation was run following [5]. For each feature set, the following steps were performed after clustering:

- Selection of top 5 most distant cluster pairs: select and remove the leaf clusters corresponding to the pair with the maximum centroid distance, repeat this for the remaining pairs, until 5 are selected.
- For each selected cluster pair (A,B), 16 triples were created (AAB, ABA, BAB, BBA), each consisting of 3 sound files: a reference sound file from one cluster of the pair, and then one sound file from each cluster of the pair. The cluster order was balanced.
- A listening test was run in which for each triple 5 listeners were asked which of the two sound files is more similar in emotion or speaking style to the reference. Their answer was counted as correct if they chose the file from the same cluster as the reference.

The listening test allows a subjective evaluation of two aspects of the clustering process. First whether the units assigned to a cluster exhibit a similar expressive state. Second whether the spread of clusters sufficiently spans the space of expressions so that there are perceptual differences between data from different distant cluster pairs. Thus, the *% correct cluster measure* in Table 1 can be viewed as a subjective form of within-to-between class ratios. It is clear that the use of *featC*, with the highest classification performance, satisfies these requirements to a greater extent than the alternative feature sets. The clustering obtained using *featC* was therefore chosen for training the ETTS models.

4.3. Training ETTS models

To train and evaluate the ETTS, the book was divided into 5 test[4] and 51 training chapters. To be confident of ensuring no mismatch between the text and audio, only units for which the lightly supervised segmentation [4] achieved 100% word accuracy against the book text were retained for training. At this threshold, 4,809 (73.4%) of the 6,554 units in the training chapters were selected. The ratio of narration, direct speech and carrier units was roughly equivalent to the full set of data with a slight change in the proportion of narration (70% to 74%) and direct speech (from 26% to 22%) units. This is due to the lightly supervised approach removing more expressive speech.

HMM training was performed using a modified version of the HMM-based speech synthesis toolkit (HTS) v2.2 [13]. The speech waveforms were sampled at 16 kHz. The observation vector consisted of the static, delta and delta-delta of 40 mel-cepstral coefficients, $logF_0$, and 21 aperiodicity bands (bark-scaled [14]). The spectrum was obtained with a pitch synchronous analysis, and the aperiodicity with PSHF [15]. The models were 5 state left-to-right multi-space probability distribution hidden semi-Markov models (MSD-HSMM) [8]. The context features were determined using a proprietary front-end text processor.

Two HMM-TTS systems were produced to represent the expressive speech (see Section 3): *decision tree* - cluster questions used in the decision tree construction; *AESS* - average expression speech synthesis. The expression-independent model trained before applying expression adaptive training for AESS was used as a *baseline*.

4.4. TTS evaluation

It was assumed that at synthesis time the appropriate cluster can be reliably chosen. To simulate this in the evaluation, the clustering was performed on both training and test set together. This way, the cluster assignment of the test sentences is known. This information was used in synthesis with the decision tree (DT) and AESS systems.

Two aspects of expressive voices need to be evaluated separately: the overall synthesis quality, and the expressiveness. Quality can be evaluated via a standard preference (paired comparison) test, in which listeners are asked which synthesised version of a sentence sounds better. This test set-up does not provide the listener with the context in which the sentence originally appeared. Hence it is important that only sentences for which the context is unlikely to be relevant are used. From the 5 test chapters, 50 narration sentences, out of those that a manual labeller had indicated as emotionally neutral, were randomly chosen.

[2] For simplicity the impact of the HSMM has been ignored for these equations.

[3] Available from librivox.org.

[4] Chapters 1, 8, 26, 28 and 39. These chapters were chosen as they contained units from a good variety of clusters during preliminary experiments.

For evaluating expressiveness, however, it is the context that determines which expression is required, or inappropriate, for a given sentence. However it is unclear how much context is needed for a listener to make such a judgement. Since it is impractical to require listeners to first read some pages of book context before judging each sentence, the human version of the sentence was used as the reference, to convey to the listeners which expression was appropriate. Listeners were asked to choose which of 2 synthesised versions of the sentence sounded more similar to the reference. For this test, 50 direct speech sentences were chosen from the 5 test chapters, which a manual labeller had indicated as emotionally non-neutral. The 100 test sentences together covered 19 out of the 20 clusters.

Baseline	Expression Rep.		No	p
	DT	AESS	preference	
39.6%	44.0%		16.4%	.1882
51.9%		42.6%	5.5%	.0503
	46.4%	41.6%	12.0%	.1708

Table 2. Synthesis quality: neutral narration preference tests.

Baseline	Expression Rep.		p
	DT	AESS	
48.6%	51.4%		.2923
29.5%		**70.5%**	.0000
	45.4%	54.6%	.0290

Table 3. Expressiveness: non-neutral direct speech similarity tests. Forced choice. Significantly (two-tailed p-test; $p < 0.025$) higher value in boldface.

Listening tests were crowd-sourced via CrowdFlower using Mechanical Turk workers located in the US [16]. Each test sentence was evaluated by 10 listeners. Tables 2 and 3 show the results. In terms of quality, there was no statistically significant difference between any of the three systems. With respect to expressiveness, however, the *AESS* was significantly more similar to the human reference speech than the baseline. This is proof of concept that the proposed unsupervised clustering and the AESS training manages to capture relevant characteristics of expressive speech. The *decision tree* system seems to fall in between the two.

5. CONCLUSION

This paper has described initial investigations into improving the expressiveness of statistical speech synthesis systems. An approach is proposed based on unsupervised clustering of audiobook data followed by HMM-TTS construction using either cluster questions in the decision tree construction or average expression speech synthesis (AESS) with cluster-based linear transform adaptation. Synthesis experiments show that the AESS built from the unsupervised clusters better reflects the expressiveness of human speech than a baseline expression-independent system.

The classification performance of the unsupervised clustering depends on the feature set. Future work will examine optimising the cluster feature set, including investigating the use of text and audio features. In this paper the appropriate expression cluster was assumed known in synthesis. Automatically determining the cluster from text is essential for applications such as ebook reading. Future work will look into approaches to learn the mapping of the clusters from the training material to arbitrary text.

6. REFERENCES

[1] M. Schröder, F. Burkhardt, and S. Krstulovic, "Synthesis of emotional speech," in *Blueprint for Affective Computing*, K. R. Scherer, T. Bänziger, and E. Roesch, Eds., pp. 222–231. Oxford University Press, 2010.

[2] Y. Zhao, D. Peng, L. Wang, M. Chu, Y. Chen, P. Yu, and J. Guo, "Constructing Stylistic Synthesis Databases from Audio Books," in *Proc. of Interspeech-ICSLP*, 2006.

[3] K. Prahallad, A. Toth, and A. Black, "Automatic Building of Synthetic Voices from Large Multi-Paragraph Speech Databases," in *Proc. of Interspeech*, 2007.

[4] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proc. of Interspeech*, 2010.

[5] E. Székely, J. Cabral, P. Cahill, and J. Carson-Berndsen, "Clustering Expressive Speech Styles in Audiobooks Using Glottal Source Parameters," in *Proc. of Interspeech*, 2011.

[6] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 Speaker State Challenge," in *Proc. of Interspeech*, 2011.

[7] 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, no. 3, pp. 503–509, 2005.

[8] 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 Algorithm," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, pp. 66–83, 2009.

[9] B. Schuller, "Recognizing Affect from Linguistic Information in 3D Continuous Space," *IEEE Trans. on Affective Computing*, vol. 2, no. 4, pp. 192–205, 2011.

[10] B. Schuller and L. Devillers, "Incremental Acoustic Valence Recognition: an Inter-Corpus Perspective on Features, Matching, and Performance in a Gating Paradigm," in *Proc. of Interspeech*, 2010.

[11] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proc. of ASRU Workshop*, 2009.

[12] D. Pelleg and A. Moore, "X–means: Extending k–means with efficient estimation of the number of clusters," in *Proc. of 17th International Conference on Machine Learning*, 2000.

[13] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, and T. Nose, "The HMM-based speech synthesis system (HTS)," `http://hts.sp.nitech.ac.jp/`.

[14] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS System for Blizzard Challenge 2010," in *Proc. the Blizzard Challenge Workshop*, 2010.

[15] P. Jackson and C. Shadle, "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech.," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, pp. 713–726, Oct. 2001.

[16] S. Buchholz and J. Latorre, "Crowdsourcing preference tests, and how to detect cheating," in *Proc. of Interspeech*, 2011.