Constructing Stylistic Synthesis Databases from Audio Books

Yong ZHAO[1], Di PENG[2], Lijuan WANG[3], Min CHU[1], Yining CHEN[1], Peng YU[1], and Jun GUO[2]

[1]Microsoft Research Asia, Beijing, China
[2]School of Information Engineering, Beijing Univ. of Posts and Telecommunications, China
[3]Department of Electrical Engineering, Tsinghua Univ. China
{yzhao, minchu, ynchen, rogeryu}@microsoft.com

Abstract

In this paper, we explore how to construct stylistic TTS databases from audio books, in which a storyteller performs multiple roles. The goal is to identify and build a set of speech corpora, each of which not only portrays a representative voice style performed by the speaker, but also has sufficient sentences to synthesize natural speech using unit selection approach. We solve the problem in two procedures: first, by representing each role with Gaussian Mixture Models (GMM), all speech data are partitioned into a number of voice style clusters with a criterion that maximizes the likelihood of all utterances with respect to roles' speaker models; then, pruning in terms of both acoustic and prosodic measures is followed to purify the clusters. The resulting 4 voice styles are subjectively interpreted as Neutral, Young, Elder and Adult, respectively. Perceptual experiments show that the proposed approach can synthesize speech with the recognizable voice styles with an average 72.5% identification rate, and the synthesized speech sounds better than those synthesized with utterances from a single role.

Index Terms: expressive speech synthesis, voice style, emotional speech

1. Introduction

Most state-of-the-art text-to-speech (TTS) systems adopt corpus-driven approaches due to their capability to generate highly natural speech. In these systems, the most suitable speech segments are selected from a very large unit inventory, and then concatenated to generate the output speech. The synthesized speech achieves natural variation by inheriting the characteristics of the source corpus, and likewise the naturalness of the speech strictly relies on the size and the coverage of the unit inventory.

For the same reason, the data-driven approach is employed to synthesize emotional speech [1][2]. In [1], several emotional speech corpora were created for each emotional state, and speech with appropriate emotions was synthesized from by switching between the emotional corpora. However, collecting a large amount of speech with the right speaking style and emotion is much more difficult than with the neutral style. Thus, it makes delivering a number of emotional voices practically inextensible.

In this paper we employ audio books as a source of emotional speech. As a naturalistic speech corpus, an audio book has many inherent properties useful for emotional synthesis [3][4]. In general, it consists of a large amount of expressive speech, which is performed by one professional speaker in a sound-proof studio. The storyteller attempts to vary his speaking style and voice quality with respect to different roles appearing in the book, in order to endow each with a representative voice and distinguish between them by means of sound only. At the same time appropriate emotions are portrayed in the roles' speech according to the plots and emotional contexts.

As a first step in studying audio-book corpora, we focus on how to identify representative voice styles in a corpus, and construct stylistic databases to synthesize speech bearing those particular voice styles. Here, voice style denotes the way of speaking in which the storyteller performs a role. Also, it is assumed that emotions inside roles are just natural variation of the role's voice.

An intuitive approach to build stylistic voices is to treat each role as a new speaker and create a separate database per role with that role's utterances. However, since roles appear quite unevenly in the book, not all roles have a sufficient number of sentences to build a high-quality voice database. Furthermore, some roles sound quite similar to each other due to sharing the same, or similar, gender, age and personalities, and hence the storyteller does not intensively discriminate them. Considering these constraints, we are not able to build separate databases for each role from an audio book.

Therefore, it becomes essential to identify a number of representative voice styles in an audio book. The key requirements for these voices are: they should sound different from each other and represent a steady speaking style in the audio book; there are enough utterances in each style for building a voice database capable of synthesis of natural speech.

Here, we solve the problem in two procedures: first, clustering speeches from all roles into several representative voices, with a criterion that maximizes the overall likelihood of the whole speech corpus with respect to roles' speaker models; then, pruning is performed in terms of both acoustic and prosodic measures to purify these stylistic databases. Perceptual experiments are conducted to evaluate the identification ratio and the naturalness of synthetic speech in these voice styles.

This paper is organized as follows. Section 2 introduces the audio-book corpus. Section 3 describes the system framework and the individual components. The experiments and the perceptual evaluation results are presented in Section 4, and conclusions in Section 5.

2. Speech corpus

For this study, we used a fictional audio book, narrated by a professional male speaker. The storyteller tells the whole story in a fully expressive way.

In order to apprehend acoustic characteristics of roles, it is necessary to associate utterances with characters [5]. First, quoted sentences are identified in the story text according to the signs of quotation marks, and labeled with the corresponding characters; then, the speech waveforms are forced-aligned to the story transcriptions with the regular speech recognition tools [6], where, as a result, we divide the speech into utterances and assign them character labels.

Up to 80 characters show up throughout the audio book, but appear considerably unevenly. As shown in Table 1, we list in descending order the top 10 characters (referred as R01, R02, …, and R10, respectively) in terms of the number of sentences (columns *Sent. Num*. and *Sent. %*). The narrator, R01, on his own, has more than half of the sentences, and the top 10 characters occupy in total 83% of the whole corpus, in contrast to the remaining 70 roles (referring to row *Others*), which together have only]17% of the utterances. Therefore, we focus on studying these top characters.

Table 1. *Attributes and acoustic characteristics of top 10 roles. (With regard to attribute Gender, M stands for male, and F for female; as for attribute Age, Y denotes young, A, middle-aged adult, and E, elder.)*

	Gender	Age	Sent. Num.	Sent. (%)	F_0 (Hz)	Duration (ms)
R01	M	A	9639	53.1	139.8 ± 15.0	82.7 ± 10.2
R02	M	Y	1363	7.5	164.1 ± 24.8	94.3 ± 20.4
R03	M	Y	903	5.0	175.8 ± 23.9	93.6 ± 22.4
R04	F	Y	832	4.6	177.9 ± 23.7	89.0 ± 17.3
R05	M	E	545	3.0	189.0 ± 21.8	104.3 ± 18.6
R06	M	A	449	2.5	136.4 ± 19.4	102.3 ± 19.7
R07	M	A	350	1.9	186.3 ± 33.6	96.6 ± 29.8
R08	M	E	344	1.9	153.9 ± 30.9	119.3 ± 26.0
R09	M	A	332	1.8	174.8 ± 20.8	89.6 ± 17.2
R10	M	A	319	1.8	158.7 ± 29.3	110.1 ± 22.9
Others	-	-	3079	17.0	-	-

Prosodic characteristics such as pitch and segmental duration are measured based on roles' utterances. First, for each prosodic feature, the average of all vowel segments in an utterance is taken as the global feature of that utterance; then the means and standard deviations (SD) are calculated on the global features of the role's utterances to describe that role's prosodic properties (see columns F_0 and *Duration* in Table 1, written as mean ± SD).

Noticeable differences between the acoustic profiles of roles confirm that the storyteller performs roles with distinctive voices. R01 has the smallest mean and SD on both F_0 and segmental duration, which imply that the narrative part is read in a fast and emotionless way, as one would expect from a narrator. To make a further analysis, we examine the relationship between prosodic properties with two basic attributes of a role, gender and age. Young roles, R02, R03, and R04, share more similar acoustic attributes than Adults and Elders. Also, on average, Young characters speak faster than Adults, and Elder characters speak the slowest among the three groups. The difference induced by gender is weak, probably because the dubbed voice is essentially from a male, or only a female character could not explain much of gender's affect.

3. System framework

The task of the system is to identify a set of representative voice styles in an audio book, and to assign to each voice style with appropriate utterances, which should reflect the characteristics of the corresponding style.

First, roles are represented by GMMs, and all speech data are partitioned into a number of voice style clusters with a criterion that maximizes the likelihood of all utterances with respect to roles' speaker models. Then, pruning in terms of both acoustic and prosodic measures is followed to purify these clusters. In result, we could synthesize speech reflecting the characteristics of these voice styles by synthesizing from the corresponding corpora. Figure 1 depicts the system flowchart.

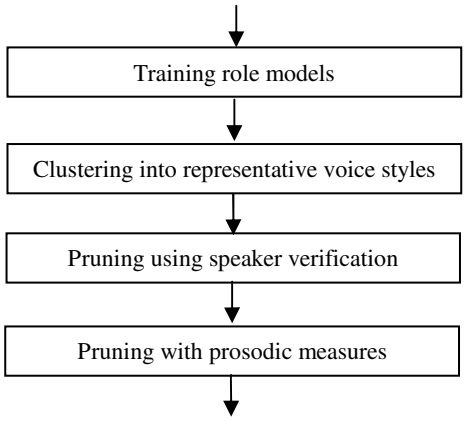

Figure 1. *Flowchart of constructing stylistic corpora.*

3.1. Training role models

By assuming every role as an individual speaker, a GMM is trained for each role. Direct Maximum Likelihood (ML) training of GMM model requires a large number of observations, which is not always available for every role. To solve this, a role independent GMM, or Universal Background Model (UBM) [7] is first trained on all utterances in the audio book, and role models are adapted with utterances from the corresponding role through Maximum A Priori (MAP) adaptation. The log-likelihood ratio between the role model and UBM model is employed as the match score of the utterance to that role.

3.2. Clustering into representative voice styles

Identifying a number of representative voice styles can be regarded as a clustering problem, since we need to combine utterances into groups which satisfy the criteria that utterances sound similar inside the groups, and different from ones from other groups. However, by reasonably assuming that the representative voice styles are always generalized from a certain role as kernel, we could obtain prior knowledge on the structure of the data. Then the problem is transformed as how to find M

roles from all roles that could partition the speech corpus in an optimal way.

Here we propose a clustering algorithm similar to K-means clustering. The criterion for optimization is to maximize the overall likelihood of the whole speech corpus with respect to M-roles' models. The objective function is defined by:

$$\sum_{j=1}^{M}\sum_{i\to r_j} LLR(X_i, r_j) \quad (1)$$

where there are M roles r_j, $j = 1,2,\cdots,M$, and $LLR()$ is the log-likelihood ratio of observation X_i from utterance i with respect to the role model of r_j, which is the nearest neighborhood to i.

Considering the number of roles in a storytelling corpus is relatively limited, we solve the problem by simply enumerating all possible combinations and choose the case maximizing the objective function.

Algorithm:

1. Initialize an M-subset from the set with N roles.
2. Associate each utterance i with one role in the M-subset closest to the utterance, i.e. the maximal likelihood on the role's speaker model. Sum up LLR of all utterances on their assigned roles.
3. Select next M-subset and repeat step 2, until exhausting all M-combinations. Choose the subset with maximum LLR sum.

The clustering based voice identification differs from k-means algorithm in that the speaker models for roles are adopted as the centroids of clusters, and centroids are constrained to move only inside a set of role models, rather than freely, to minimize the overall distortion function. As a consequence, it provides a natural way to incorporate prior knowledge of data structure.

3.3. Pruning using speaker verification

Not all utterances inside the same cluster sound similar enough to each other, since some utterances sound far from all clusters, and have to be assigned to the nearest group. Thus, we verify the utterance in terms of both acoustic and prosodic measures to purify these stylistic databases.

As for acoustic confidence measure, we still employ LLR score of the utterances with respect to the role model, just like speaker verification does. The utterance is accepted as from a voice if its score is above a preset threshold for the speaker model of that kernel role; otherwise, it will be rejected.

3.4. Pruning with prosodic measures

GMM based role models do not account for temporal information of speech waveforms, and lack the ability to capture speaker's prosodic characteristics. Therefore, we continue to employ prosodic confidence measures to remove the utterances that sound rather different in prosody from the voice's normal speaking way. Here, we measure the deviation of an utterance from its expected prosody contour that is predicted by prosodic decision trees with the sentence transcription as input.

Similar to what has been done in [8], Classification And Regression Tree (CART) is trained on all samples in the voice corpus to predict prosodic parameters of segments confined with certain contexts. The parameters consist of F_0, dynamic range of F_0, and segmental duration. The set of splitting questions relates to phonetic and prosodic context features such as left and right phone context, position in phrase, and position in word.

First, the mean and standard deviation are calculated over samples within CART leaves, and regarded as the distribution for segments sharing the same context. Based on it, the deviation of a segment is measured by z-score normalization. The prosodic deviation of an utterance is just the average deviation over all its segments.

4. Experimental results

We tested the system with the audio-book corpus. UBM was trained from all utterances in the audio book, and speaker models of the top 10 roles were obtained separately by adapting with 250 utterances of each role. Acoustic features consist of 13-order MFCCs and pitch, and their delta coefficients. Only means were adjusted during the adaptation.

4.1. Analysis of clustered voice styles

First, we evaluated the effect of the clustering algorithm. By setting the number of clusters to 4, we clustered the speech of all roles into 4 voice styles, which grow from kernel roles R01, R01, R02, R08, and R09, respectively. Table 2 lists the utterance number of each clustered voice style and its distribution across roles. For each role, the shadowed cells indicate which voice style occupies the maximal proportion of the role's utterances. It is rational to believe that the role speaks similar to its shadowed voice style, where the proportion can be looked upon as an index of the similarity.

By comparing the shadowed cells with the age of roles, it is found that there exists a rough relationship between the cluster and the age: cluster R01 almost overlaps with role R01; cluster R03 includes nearly exclusively three teenagers, R02, R03, and R04; cluster R08 relates more closely to the elder roles, R05 and R08; cluster R09 include most of utterances from middle-aged adults, R06, R09 and R10.

Table 2. *Utterance distribution of the identified voice styles.*

	R01	R02	R03	R04	R05	R06	R07	R08	R09	R10	Others	Sum
Age	A	Y	Y	Y	E	A	A	E	A	A	-	-
Sent. Num.	9639	1363	903	832	545	449	350	344	332	319	3079	18155
Neutral (R01)	9359	60	1	3	0	43	0	0	1	0	264	9731
Young (R03)	124	1122	807	672	142	79	189	0	0	91	1451	4677
Elder (R08)	82	94	24	145	352	144	86	335	0	97	677	2036
Adult (R09)	74	87	71	12	51	183	69	15	331	131	687	1711

Therefore, we interpreted these clusters with explicit style names, R01 as Neutral, R03 as Young, R08 as Elder and R09 as Adult. Furthermore, we perceived these voice styles with particular characteristic descriptions: style Neutral represents the flat, less emotive speaking style; style Young represents the higher pitch, faster rate and younger voice; style Elder exhibits the lower pitch, slower and older voice, and style Adult, the middle-aged adult with faster, emotional and decisive voice.

Though the table shows that most of roles have evident affiliations to one style, some roles disperse their speech into multiple styles, like R06 and R10. This is mainly because these roles do not match any of the resulting voice styles. This can also be proved by the fact that when setting the number of clusters to 5, R10 is separated as a new voice style in addition to the existing 4 styles.

Pruning was conducted on the resulting clusters. As for the utterance verification module, we set the pruning ratio for style Neutral as 10%, and others 40%, since the style Neutral contains mainly the narrator's utterances, and so is already essentially pure.. In pruning with prosodic measures, 5% of the utterances for each style are removed. It is perceptually observed that these utterances sound somewhat emotionally intensive, or with a salient prosody.

4.2. Identification of voice styles

We built four stylistic voices with our Mulan TTS system [9]. A series of perceptual evaluations were conducted on the synthesized waveforms to evaluate the performance of the proposed method.

First, we tested the identification of the claimed styles, where subjects were asked to select a style type, which best matches the characteristics of the played utterance. To help subjects learn the styles, each style is annotated with a few descriptive words summarizing the voice' characteristics, as shown in Table 3. Also, before the experiment, participants will listen to 5 utterances (original recordings from the stylistic corpus) per style to get acquainted with these styles. The experiment consisted of 10 sentences synthesized with 4 styles separately, totally 40 stimuli, which are played in a random sequence. 8 subjects participated in the test.

The identification rates are presented in Table 3. The results show that the claimed voice styles are recognized with a high rate of accuracy, with Neutral: 80.0%, Young: 60.0%; Elder: 88.3% and Adult: 61.7%. Styles Young and Adult are shown to be confusing pairs, as is attributed to their close pitch range and speaking rate.

Table 3. *Identified styles from four stylistic databases.*

Voice style	Style description	Subjects' response (%)			
		Neutral	Young	Elder	Adult
Neutral	Narrative, emotionless	80.0	10.0	3.3	6.7
Young	Fast, high-pitched	1.7	60.0	6.7	31.7
Elder	Slow, low-pitched	0.0	5.0	88.3	6.7
Adult	Fast, decisive	1.7	25.0	11.7	61.7

4.3. Evaluation of naturalness

To evaluate the naturalness of the synthesized speech, a preference test was conducted. The reference system was the direct approach that builds a voice database with only utterances from the kernel role corresponding to a voice style.

The experiment included 15 sentences for each voice style, 60 utterance pairs in total. 8 subjects participated in the test and they were forced to choose from each pair one that sounded more natural. The result for the preference test is given in Table 4. It shows that the synthetic speech obtained with the proposed database construction algorithm sounds better on average than that with the direct approach. This illustrates that if we can find more speech waveforms homogeneous to the claimed style, the synthesized speech quality would be improved. This point is especially important for style Adult and Elder, since their naturalness has been significantly improved due to expansion of the database.

Table 4. *Preference test results.*

Voice style	Baseline (%)	Proposed (%)
Neutral	52.2	47.8
Young	41.1	58.9
Adult	35.6	64.4
Elder	34.4	65.6

5. Conclusion and discussion

In this paper, we have proposed an approach to construct stylistic databases from an audio book using two procedures: first clustering speeches from all roles into several representative voices, and then pruning the clusters in terms of acoustic and prosodic measures. The clustered 4 voice styles were subjectively interpreted with explicit style assignments, Neutral, Young, Elder and Adult. A perceptual experiment using our Mulan TTS system showed the synthesized speech is recognized at a high agreement to the claimed voice styles, and the synthesized speech sounds better than those synthesized with the direct approach.

Further work involves discriminating various emotions inside the voice styles. Also, we will investigate the capability of the model-based synthesis on these style sets, due to its robustness to wild sample units and the advantage in limited data footprint.

6. References

[1] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, A corpus-based speech synthesis system with emotion. Speech Communication, 40, 161-187, 2003.

[2] A. W. Black, "Unit selection and emotional speech", in *Proc. of Eurospeech 2003*, Geneva, 2003.

[3] C. O. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling", in *Proc. of Interspeech 2005*, Lisbon, 2005.

[4] L. Wang, Y. Zhao, M. Chu, etc. "Exploring expressive speech space in an audio-book", in *Proc. of Speech Prosody 2006*, in press.

[5] J. Y. Zhang, A. W. Black, and R. Sproat, "Identifying speakers in children's stories for speech synthesis", in *Proc. of Eurospeech 2003*, Geneva, 2003.

[6] J. Odell, D. Ollason, P. Woodland, S. Young and J. Jansen, *The HTK Book for HTK V3.0*, Cambridge University Press, Cambridge, 2001.

[7] D. A. Reynolds, T. F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models", Digital Signal Process, vol. 10, no. 1–3, pp. 19–41, 2000.

[8] Y. Zhao, M. Chu, H. Peng and E. Chang, "Custom-tailoring TTS voice font – keeping the naturalness when reducing database size", in *Proc. of Eurospeech 2003*, Geneva, 2003.

[9] M. Chu, H. Peng, Y. Zhao, Z. Niu and E. Chang, "Microsoft Mulan - a bilingual TTS system", in *Proc. of ICASSP 2003*, Hong Kong, 2003.