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

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

Predicting Character-Appropriate Voices for a TTS-based Storyteller System

Erica Greene[1], *Taniya Mishra*[2], *Patrick Haffner*[2], *Alistair Conkie*[2]

[1]Computer Science Department, The University of Southern California, Los Angeles, CA, USA
[2]AT&T Research, Florham Park, NJ, USA

ericargr@usc.edu, {taniya,haffner,adc}@research.att.com

Abstract

Using distinct and appropriate synthetic voices to voice the characters in a children's story would make a TTS-based digital storyteller system more engaging and entertaining, as well increase listener's comprehension of the story. However, automatically predicting appropriate voices for storybook characters is both a non-trivial and largely unexplored problem.

In this paper, we present a data-driven approach for predicting the most appropriate voices for characters in children's stories based on salient character attributes. We use Mechanical Turk to identify the character attributes that are most salient in evoking the listeners' perception that a specific character should have a particular voice, and to label the voices in our collection with attribute tags. We model the attribute-to-voice relationship with Naive Bayes. The resulting system performs significantly above chance in an objective evaluation, demonstrating the viability of our approach.

Index Terms: Speech synthesis, TTS, expressive speech, child-directed speech applications.

1. Introduction

When e-books equipped with text-to-speech (TTS) capabilities were initially launched a few years ago, there was both great excitement for and in opposition to the technology. Since then, both of these feelings have significantly lessened. The reason, quite simply, is that the TTS features are not being heavily used: automatic narration of stories using the general purpose TTS synthesizers currently embedded in these TTS-enabled ebook readers fail to live up to users' standards. The listening experience is nowhere near as engaging as listening to the human storyteller.

For a digital storytelling system to be comparably engaging as human storytellers, the system has to narrate the story using a "storytelling speech style " [1], express basic emotions such as sadness, fear, happiness, anger and surprise [2, 3], generate suspense and anticipation, recognize different characters in the stories and their personal attributes [4, 5, 6], and render the dialogs uttered by each of the characters using different character-appropriate voices. Several of these interesting aspects of digital storytelling have been addressed in the cited works. But the last mentioned feature, rendering dialogs uttered by each of the characters using different character-appropriate voices, is yet to be addressed. This was one of the goals of the digital storyteller system ESPER [7], but there does not appear to have been much work done on this topic.

In Cabral et al. 2006 [8], it has been shown that selecting an appropriate voice for a character in a digital storyteller system is significant for understanding a story, perception of the emotional content, credibility of the voice, and listener satisfaction with the voice. Using character-appropriate voices for rendering quoted material by the different characters in a story would be particularly useful in child-directed applications. Children's stories often contain multiple characters, each with different personalities, genders, age, ethnicities, etc. Several of these characters may be anthropomorphic. When this type of rich narrative material is synthesized such that each character has its own distinct, character-appropriate voice, the outcome would be much more engaging and entertaining for kids (and adults) than using a single preselected TTS voice. Such technology also has the potential to improve story comprehension by young listeners who sometimes struggle to understand dialogue between multiple characters.

At its core, automatic storytelling using character-appropriate voices is a prediction problem: given a particular character and its attributes, which is the most appropriate voice to use to render its dialogues? Which voice *best personifies* this character? We present a first approach towards solving this problem. We focus specifically on children's stories.

Our approach consists of estimating the types of voices needed to cover the prototypical characters found in children stories (2.1), identifying the character attributes that are most salient in evoking the listeners' perception that a specific story character should have a particular voice (2.2), modeling the relationship between the character attributes and the voices (3), and evaluating the performance of our system (3.1).

2. Data and Methods

2.1. The Voices

A quick analysis of common children's stories reveals that many of the principal characters are strikingly similar: consider Gretel and Goldilocks, Cruella de Vil and Ursula, or Aladdin and Robin Hood. For our storyteller system to be successful, it must include representative voices from each of these reoccurring classes of characters.

To estimate the types of voices needed to cover the core characters, we extracted a list of characters from a collection of children's stories on Project Gutenberg. We tried to match each of these characters with at least one of the AT&T TTS voices and found our collection of voices had poor coverage of young, evil and silly voices. We filled in the sparsity with 18 voices from other publicly available sources. The composition of our final voice set is shown in Table 1.

2.2. The Attributes

When completed, the final storyteller system will select a voice based on a set of character attributes that have been extracted from the text. We wanted to use the character attributes that are most salient in evoking the listeners' perception that a particular

	Total	Male	Female
Prototype voices	7	2	5
Public AT&T	6	3	3
Commercial voices	18	10	8
Total	31	15	16

Underrepresented Categories

British	young	old	silly	evil
5	4	3	3	3

Table 1: *Set of TTS Voices.*

character should be paired with a particular voice. The attribute set should be discriminative enough to uniquely identify a voice, while compact enough to model using a relatively small amount of annotated data. Certain attributes such as age and sex are obvious choices to include, but two dimensions are insufficient to characterize the 31 TTS voices.

Instead of choosing the attributes based on the authors' subjective idea of which attributes are important, we utilized Amazon Mechanical Turk (MT) [9]. The Mechanical Turk task asked participants to listen to audio clips of voice actors voicing characters from children's stories. The turkers were then asked to describe the voice using a mixture of multiple choice and free response questions. We obtained the audio clips from children's movies and tv shows that were on www.behindthevoiceactors.com, a freely available database of voice actor demos. The resulting MT task included 60 clips from children's shows, carefully chosen to avoid categories such as anime and comic books whose characters do not appear in children's stories.

A screen shot of this attribute collection MT task can be seen in Figure 1.

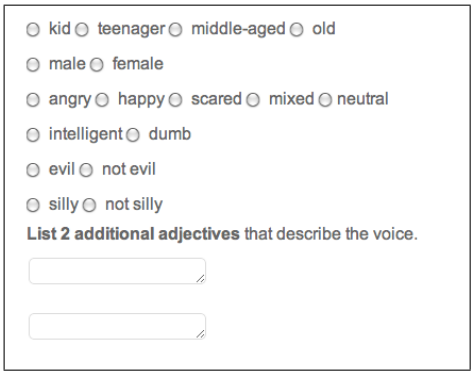


Figure 1: The first Mechanical Turk task asked participants to listen to a clip of a voice actor voicing a children's character and describe the voice.

We ran the task 5 times for each audio clip, resulting in 300 samples. All submissions that did not correctly identify the gender of the character were rejected.

We used the free-response adjectives from the MT task to determine a set of salient attributes to describe voices for children's book characters. While the free-response attribute set was diverse and generally descriptive of the voice in the clip, many of the adjectives, such as "confused" and "excited", described the emotion of the character rather than the quality of the character's voice.

We augmented the list with more appropriate attributes by parsing voice actor's profiles from a digital voice actor directory. Below are examples of the descriptions from the directory.

* I have a friendly, child-like voice that is very sweet and inviting and am very articulate. My voice ranges from a child, to preteen and adult voices. My voice qualities are warm & friendly, sultry, corporate.
* I am the 20s/30s woman with a full-bodied voice, luscious and silky yet crisp and clear. Voice ranges from light and upbeat to silky smooth. My voice sounds relatively young, warm and friendly, [and] well articulated.

We used Wordnet to identify the adjectives in the profiles, and then cluster all of the adjectives from both the MT task and the voice actor directory into roughly 20 similarity groups. After manually discarding irrelevant clusters, we chose one representative adjective from each clusters to appear in the final attribute list. Below are some of the attribute clusters with the characteristic adjective shown in bold.

* straight proficient honest **serious**
* hoarse gruff **raspy** scratchy rough grating
* giddy goofy **silly** wacky

With the exception of sex and age, we map each attribute to an integer-valued feature between 1 and 5, with 1 corresponding to 100% value 1 and 5 corresponding to 100% value 2. We restrict sex and age to take two and four values respectively. The final most salient attribute set is described in Table 2.

Name	**Values**	
Sex	male	female
Age	kid teen	middle-aged old
Intelligence	intelligent	dumb
Evilness	evil	not evil
Silliness	silly	serious
Gentleness	gentle	rough
British-ness	British	not British
Shyness	shy	confident
Naturalness	artificial	natural
Smoothness	raspy	smooth
Pitch	deep	high pitched

Table 2: *Salient Attribute Set*

In the future, we would like to use the data from the multiple choice questions in the MT task to analyze the voices from the audio clips in order to learn how to better distinguish a rich set of acoustic prosodic characteristics of the voices, such as happiness and silliness.

2.3. Data Labeling

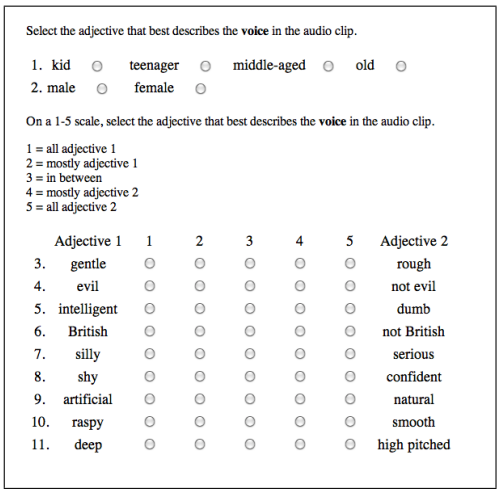
We also used Mechanical Turk to annotate the synthetic TTS voices in our collection with one or more of the salient attributes that were identified in Section 2.2.

This task provided the user with an audio clip of a single TTS voice rendering a short piece of text. The Turker was asked to rate the voice in the clip with respect to each attribute. A screen shot of the task is shown in Figure 2.

The selection of the text excerpts had a few subtle challenges. Ideally, labelers would annotate the attributes of the voice without being swayed by the content of the text being spoken. However, there is often a strong correlation between a

character's attributes and the content of the character's dialog (e.g., an evil character is likely to say mean things), and hence a common scenario that the storyteller system will encounter.

In order to control for the content of the audio clip, we selected paired text excerpts for each voice to render. Each excerpt pair contained one excerpt that provided some clues to the personal attributes of the character, and another excerpt that was character neutral.



Select the adjective that best describes the **voice** in the audio clip.

1. kid ○ teenager ○ middle-aged ○ old ○
2. male ○ female ○

On a 1-5 scale, select the adjective that best describes the **voice** in the audio clip.

1 = all adjective 1
2 = mostly adjective 1
3 = in between
4 = mostly adjective 2
5 = all adjective 2

	Adjective 1	1	2	3	4	5	Adjective 2
3.	gentle	○	○	○	○	○	rough
4.	evil	○	○	○	○	○	not evil
5.	intelligent	○	○	○	○	○	dumb
6.	British	○	○	○	○	○	not British
7.	silly	○	○	○	○	○	serious
8.	shy	○	○	○	○	○	confident
9.	artificial	○	○	○	○	○	natural
10.	raspy	○	○	○	○	○	smooth
11.	deep	○	○	○	○	○	high pitched

Figure 2: Turking Task 2

Here is an example excerpt pair:

Little Red Riding Hood (Neutral) "You go down this path to the mill and then turn to the right."

Little Red Riding Hood (Biased) "Yes, mother, you know I always like to visit dear grandmamma."

We used 40 such excerpts from the dialogs of 20 different children's characters. We created a MT task for every excerpt-voice combination and repeated each task twice, resulting in approximately $2,500$ samples of annotated audio clips.

The covariance between the attributes is shown in Figure 3, where positive and negative values are shown in red and blue respectively. The last column corresponds to whether the text rendered in the audio clip of the sample was biased or neutral. As we can see, the content of text had little effect on the value of the other attributes. While some of the relationships are relics of the quality of the TTS voices, such as the negative relationship between British-ness and naturalness, some trends such as the negative relationship between evilness and gentleness reinforce our intuition about the attributes.

3. Modeling the Attribute-to-Voice Relationship

Ideally, the choice of the voice given the attributes should be a classification problem. However, the training data as labeled in the previous section leads more naturally to the estimation of statistics such as the probability of an attribute given a voice, averaged over the context and the subjectivity of the turker. In contrast, the test conditions described in the next section, suggest a ranking problem because there is not an objective top choice. We found that the Naive Bayes classifier provided a straightforward bridge between these very different training and testing conditions. Despite their simplicity, such classifiers have

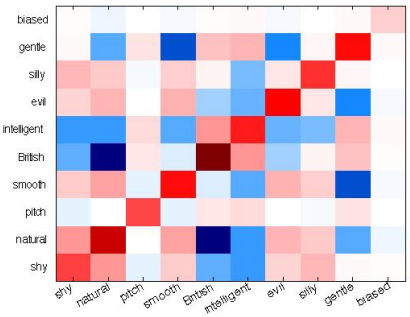


Figure 3: *The covariance matrix of attributes. Intensity indicated larger absolute value, and red and blue corresponds to positive and negative values respectively.*

been shown to be very efficient for ranking problems such as Spam mitigation. They also handle well the fact that the class prior we observe on the training data is not representative of the testing conditions as Nave Bayes priors can be set independently.

The naive bayes formulation is as follows: let $\mathcal{A} = \{a_1, \ldots, a_{11}\}$ be the set of all possible attributes. Consider each character attribute, such as evilness, silliness or pitch, as a dimension of the voice. The value of the attribute corresponds to a point on that scale. With the exception of age and sex, each a_i can have an integer value between 1 and 5. For example, a_5 is degree of silliness, where $a_5 = 1$ means very silly and $a_5 = 5$ means very serious. A user inputs a subset of the 11 possible attributes along with a corresponding value for each attribute.

For a set of attributes $\{a_1, \ldots, a_k\} \subseteq \mathcal{A}$ and a set of voices $\mathcal{V}$, the Naive Bayes decision rule is this:

$$\hat{v} = \arg\max_{v \in \mathcal{V}} P(a_1, \ldots, a_k \mid v)$$

where

$$P(a_1, \ldots, a_k \mid v) = \prod_i P(a_i \mid v)$$

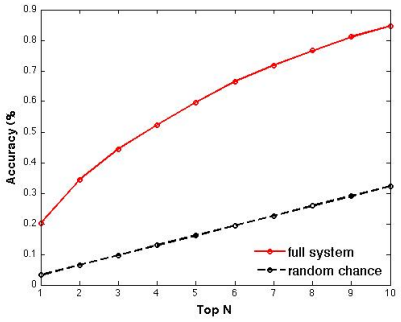
$$P(a_i \mid v) = \frac{\text{count}(a_i \text{ and } v)}{\text{count}(v)}$$

The system assumes the prior $P(v)$ is the same for all voices. Applying Laplacian smoothing to the probabilities did not significantly improve results with respect to our evaluation metrics.

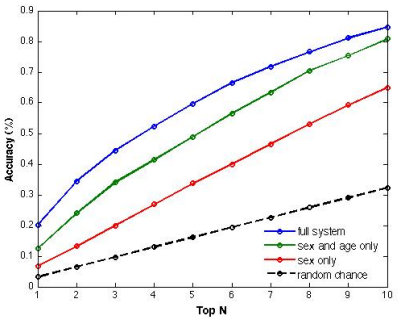
3.1. Evaluation

We evaluate our system by training our model on a subset of the data, and then predicting the voices of the held-out samples.

Each entry of the data is a tuple $(\mathbf{a}, \mathbf{v})$ where $\mathbf{a} \subseteq \mathcal{A}$ is a vector of attributes and $v \in \mathcal{V}$ is a voice. We randomly split the data into 80% train and 20% test. For each sample $\mathbf{a}$ in the test set, the Naive Bayes model returns a ranking of the voices, from most likely to least likely. The $n-$best metric returns 1 if the correct voice is one of the top n voices returned by the model. The results of this evaluation are shown in Figure 4 (a), for n ranging from 1-10. The results are averaged over 10 random splits of the data. The dashed black line is the probability of choosing the correct voice by random chance, or $\frac{n}{31}$. As we can see, the system has over a 50% chance of returning the correct voice as one of its top 5 guesses and a 85% chance of returning the correct voice in the top 10 guesses.



(a) Size of top n-list versus accuracy.



(b) Contribution of sex and age to n-best accuracy.

Figure 4

Figure 4 (b) shows the contribution of the sex and age attributes. Sex accounts for about half of the improvement over random guessing, which makes sense because our set of voices is roughly half male and half female.

A procedural drawback of this evaluation method is that we specify a value for every attribute. In practice, characters will typically have only a handful of distinguishing characteristics. The evaluation method does not take into account the information implied by the exclusion of certain attributes. We can compensate for this problem by only using the most salient attributes for each voice, a heuristic that simply disregards all attributes with a value of 3. This technique improves 1-best accuracy by 1-3 percentage points.

4. Conclusion and Future Directions

The main goal of this work was predicting the most appropriate voice for each character in a children's story in a TTS-based digital storytelling system. We took a data-driven approach towards solving this problem. Using Amazon Mechanical Turk, we learned which character attributes were most salient in evoking the listeners' perception that a specific story character should have a particular voice. Having identified the salient attributes, we used MT to annotate our collection of voices with multi-valued attribute tags. Then, we modeled the relationship between the character attributes and the voices using Naive Bayes.

For a given attribute set, our model returns a ranked list of the voices, from the most appropriate to the least appropriate. To evaluate our system, we repeatedly tested it on a randomly selected 20% of the data, while training on the remaining 80%. Aggregated evaluation results showed that our system's predictions of character-appropriate voices aligned with the human listeners significantly more often than chance.

A limitation of our current model is that we assume a uniform prior over the voices. The true prior distribution for a voice is the probability over all characters in all children's books that there is a character that is best suited for that voice. Computing this probability directly is infeasible, but we should not simple disregard the priors because they can have a significant influence on the classification system. One possible method to determine the priors is to assume the prior distribution has the maximum entropy given some constraints about the data. We can use constraints drawn from the ranking data to adjust the distribution.

So far, we have only used an objective test to evaluate our approach. In the future we hope to do a perceptual study to measure appropriateness of the automatically selected voice as perceived by human subjects of different age-groups, genders and ethnicities. In this study, each individual stimulus will contain longer multi-character exchanges, where each character's voice is selected by our system. The human subjects will be asked to rate the appropriateness of each character voice individually, and of the group as a whole. Such an evaluation is motivated by the fact that the system predicted voice for each character has to not only be distinctive and appropriate for the character in isolation, but it must also make sense in an ensemble of various characters.

5. References

[1] Theune, M., Meijs, K., Heylen, D., and Ordelman, R., "Generating expressive speech for storytelling applications," in IEEE Transactions on Audio, Speech, and Language Processing 14 (4), 2006, pp. 1137-1144.

[2] Alm, C., and Sproat, R., "Perceptions of emotions in expressive storytelling," in Proceedings of InterSpeech 2005, Lisbon, Portugal, Sep. 4-8, 2005

[3] Alm, C., Roth, D., and Sproat, R., "Emotions from text: machine learning for text-based emotion prediction," in Proceedings of HLT/EMNLP 2005, October 6-8, 2005, Vancouver, Canada.

[4] Mamede, N. Chaleira, P., "Character Identification in Children Stories," in the Lecture Notes in Computer Science, ISSU 3230, 2004, Springer-Verlag, pp. 82–90.

[5] Elson, D. K., and McKeown, K. R., "Automatic Attribution of Quoted Speech in Literary Narrative,", in Proceedings of AAAI 2010, Atlanta, Georgia, USA, July 1115, 2010.

[6] Glass, K., and Bangay, S., "A naïve salience-based method for speaker identification in fiction books," in Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa, Pietermaritzburg, South Africa, 2007, pp. 1–6.

[7] Zhang, J., Black, A., and Sproat, R., "Identifying speakers in childrens stories for speech synthesis," in Proceedings of EUROSPEECH 2003, September 1-4, 2003, Geneva, Switzerland.

[8] Cabral, J., L., Oliveira, Raimundo, G., and Paiva, A., "What voice do we expect from a synthetic character?" in Proceedings of SPECOM, June 25-29, 2006, St. Petersburg, Russia, pp. 536-539.

[9] Amazon mechanical turk. http://www.mturk.com/