The NTUT Blizzard Challenge 2012 Entry

Yuan-Fu Liao, Chia-Chi Lin and Jiun-Yan Pan

Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan

yfliao@ntut.edu.tw

Abstract

This paper describes our HMM-based speech synthesis system (HTS) [1] submitted to Blizzard Challenge 2012 [2]. This is our first English TTS and also our first audiobook application. In this system, not only linguistic but also semantic features beyond sentence level are extracted including the (1) semantic topics and (2) punctuation marks (PMs) of current and surrounding sentences and (3) number and forward and backward positions of sentences in a paragraph. Especially, Latent Dirichlet Allocation (LDA) [3]-based approach was adopted to analyze the topic of an input sentence and applied to both (1) decision tree-based clustering and (2) adjust the durations of inter-sentence breaks.

Index Terms: speech synthesis, HMM, HTS, audiobook

1 Introduction

This paper describes our HTS-based speech synthesis system submitted to Blizzard Challenge 2012, the open evaluation that compares the performance of different TTS systems with a common speech database. This is not only our first English TTS but also our first audiobook system.

Usually, the current TTS systems treat each input sentence equally and synthesize speech in a sentence by sentence way. In other words, the relationship between sentences in a paragraph or a discourse is often not taken into account. The consequence is that every synthesized utterance may sound similar in speaking style and is not good for long-time listening. For example, a listener may feel the synthesized speech is boring and is hard to follow. This concern is especially true for audiobook applications.

In order to build a voice suitable for audiobook applications, relationship between sentences should be well considered and every sentence should be adjusted in different way. So, in this paper, we try in the first time to extract both linguistic and semantic features beyond sentence level. The basic idea is that:

(1) A TTS system should be aware of the topics of current and surrounding sentences and give different topic different speaking style. Moreover, if there is a topic switching between two sentences, the duration of synthesized break between these two sentences should be longer than those without inter-sentence topic switching. In this way, a listener may perceive the change of topics easier.

(2) PMs give the cues about the different relationship between sentences. So the PMs of current and surrounding sentences should be known by a TTS system.

(3) The number of sentences and the forward and backward position of a sentence in a paragraph or even a discourse should also be given to a TTS system.

We therefore adopt a latent topic model [3] and a natural language processing (NLP) parser [4] to build many higher-level cues related to inter-sentence relationship into our TTS system.

Unfortunately, during the implementation of those ideas into our system, some serious mistakes were made in preparing the label files. So almost all contextual information (except 5-gram phone context) was ignored by the decision tree-based clustering procedure. Therefore, our result submitted to Blizzard Challenge 2012 is not correct and is not good enough. Since, we don't have enough time yet to correct those errors (still ongoing). We will not discuss the official evaluation results here, but only describe the (1) block diagram of our system, (2) voice building procedure and (3) give some LDA experimental results in this paper.

2 HTS-based English Audiobook TTS

In the following sub-section, several sub-modules of our system will be described in more detail including (1) linguistic and semantic cues extraction frontend, (2) inter-sentence break prediction and (3) question set above sentence level.

2.1 Linguistic and Semantic Cues Extraction Frontend

Figure 1 shows the block diagram of our linguistic and semantic extraction frontend. It has five modules including (1) a NLP parser (text normalization and part of speech (POS) tagging), (2) a letter to sound, (3) a latent topic model, (4) phone aligner and (5) inter-sentence break prediction. The outputs of the frontend are the HTS label files feed into decision tree clustering.

In our system, Stanfold NLP parser [4] and topic modeling toolbox (TMT) [3] are adopted. The former is applied to generate POS of each word. The latter is based on LDA method and is trained to analyze input text to get the topic each sentence, respectively.

On the other hand, we directly used the letter to sound, phone segmentation information and confidence scores (generated by an automatic speech recognizer) released along with the database [2].

Finally, the extracted topic tags are used to adjust the duration of inter-sentence breaks and all information was then integrated to generate HTS label files for all sentences.

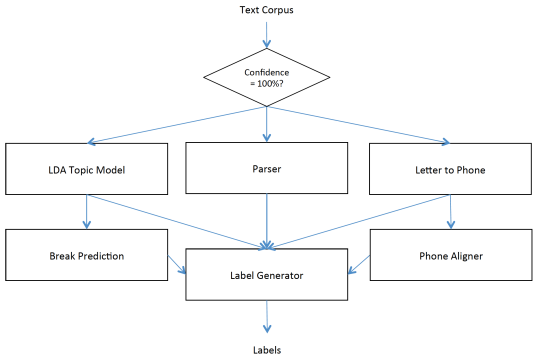

Figure 1: The block diagram of the linguistic and semantic cues extraction frontend.

2.2 Inter-Sentence Break Prediction

Figure 2 shows the block diagram of the topic-switching-based break duration prediction module. The LDA-based topic model was trained using all text material in the database. Then the break prediction was done using a decision tree-based approach.

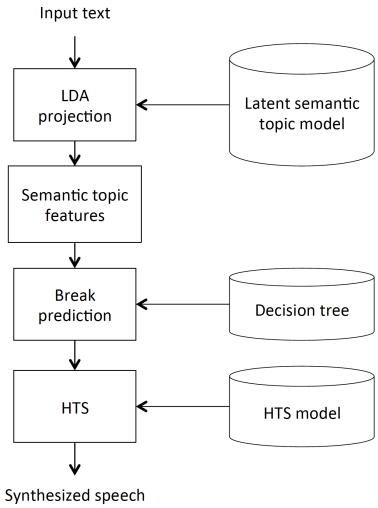

Figure 2: The block diagram of the inter-sentence break prediction module.

2.3 Question Set for Clustering

The question set used for clustering all the context-dependent phones is composed of 5 layers and listed in Table 1. It is worth noting that for audiobook tasks, linguistic and semantic features beyond sentence level were added including (1) the semantic topics, (2) PMs of current and surrounding sentences and (3) the number and forward and backward positions of sentences in a paragraph.

In our system, 5 different PM groups including period, comma, question exclamation and colon were considered. Moreover, a latent semantic topic model with 30 semantic topics automatically extracted using Stanfold TMT tool were adopted to analyze an input text.

Layer	Question
Phone	the names and types of current and surrounding phones (5-gram); the number and forward and backward position of a phone in a syllable
Syllable	the number and forward and backward position of a syllable in a word; is it a stressed syllable or not
Word	the part-of-speech (POS) of current and surrounding words; the number and forward and backward position of a word in a phrase
Phrase	the number and forward and backward position of a phrase in an sentence
Sentence	the punctuation mark (PM) and semantic topic of current and surrounding sentences; the number and forward and backward position of a sentence in a paragraph

Table 1: Hierarchical structure of question set for decision tree-based context-dependent phone model clustering.

3 Voice Building Settings

Here we give some detail about our voice building settings including (1) the audiobook database and waveform pre-processing, (2) sentence selection, (3) X-SAMPA phoneme set, (4) speech signal representation and (5) voice building procedure.

3.1 Audiobook Database and Waveform Pre-Processing

Four audiobooks including (1) “A Tramp Abroad”, (2) “Life on the Mississippi”, (3) “The Adventures of Tom Sawyer” and (4) “The Man that Corrupted Hadleyburg” were released this year. These books were written by Mark Twain and produced by a single speaker, John Greenman. There are in total 27,320 utterances (about 54.8 hours) in this database.

Thsee four books were pre-processed by Toshiba and came with word transcriptions, phone and syllable segmentations, stressed or unstressed syllable flags and POS tags. Short pause and silence segments are also detected from speech signal. Moreover, confidence scores of all utterances generated by an automatic speech recognizer are also given.

It is worth noting that the recording condition of the audiobook “Life on the Mississippi” is quite different with other three books. Therefore, in our system, a de-emphasis IIR filter was applied using SOX [5] software to all the recordings in this book to alleviate this problem.

3.2 Sentence Selection

The training material was basically selected by choosing only those utterances with 100% confidence scores. Besides, some texts that could be correctly normalized and handled by the Stanfold NLP parser (most of them are number sequences) were also ignored. In the end, only 14,654 utterances were used as the training data.

3.3 X-SAMPA Phoneme Set

Following the transcription data given by Toshiba, 55 phonemes encoded using the Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) were chosen as the basic synthesis units. Table 2 shows a mapping table between Alphabet and X-SAMPA.

For each synthesis unit, a HMM with 5 states, left-to-right transition and diagonal covariance matrix is adopted.

Alphabet	ao	aa	iy	uw	eh	ih	uh	ah
X-SAMPA	O:	A:	i:	u:	E	I	U	V
Alphabet	ax	ae	ey	ay	ow	aw	oy	er
X-SAMPA	@	{	eI	aI	ou	aU	OI	3`
Alphabet	axr	ehr	uhr	aor	aar	ihr	iyr	awr
X-SAMPA	@`	Er	Ur	Or	Ar	Ir	Ir	aUr
Alphabet	p	b	t	d	k	g	ch	jh
X-SAMPA	p	b	t	d	k	g	tS	dZ
Alphabet	f	v	th	dh	s	z	sh	zh
X-SAMPA	f	v	T	D	s	Z	S	Z
Alphabet	hh	hv	m	em	n	nx	en	ng
X-SAMPA	h	h_v	m	m=	n	4~	n=	N
Alphabet	l	el	r	R	dx	y	w	q
X-SAMPA	l	l=	r	r`	4	j	w	?

Table 2: The mapping table of X-SAMPA and Alphabet phoneme set for English voice.

3.4 Speech Representation

All speech data were first up-sampled from 44.1 Khz to 48 kHz. Then 34-order mel-generalized cepstrum (MGC) [6] and fundamental frequency, F0 were extracted using A Robust Algorithm for Pitch Tracking (RAPT) [7] algorithm as the spectral and excitation parameters (with 5ms frame shift). Besides, their first and second order derivative features were also generated to form a 105-dimentional feature vector for each speech frame.

3.5 Training Procedures

The voice building steps are showed in Fig. 3. Here Deterministic Annealing Expectation and Maximization (DAEM) [8] and Minimum Generation Error (MGE) [9] training algorithm supported by HTS version 2.2 were utilized in order to build a better voice. The numbers of iterations for DAEM and MGE are experimentally set to 10 and 50, respectively.

These two procedures are in fact very time-consuming. But according to our preliminary experiments, they do improve the voice quality a lot. So we still applied them here.

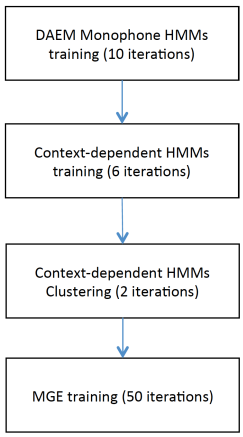

Figure 3: The block diagram of the voice building procedure using HTS version 2.2.

4 LDA Experimental Results

First of all, Table 3 shows the extracted 30 latent topics with their corresponding 10 most important words. It may not be easy to understand the found latent semantic topics. However, it seems that topic 4 is related to the landscape around Mississippi river and topic 12 is talking about numbers and units. Especially, almost all German words were grouped into topic 15. So the results may be reasonable and could be used to generate topic tag for each sentence well.

5 Conclusions and Future Works

This paper describes our first English TTS and also our first audiobook application. From the evaluation results, it seems there is still a lot of room for improvement. In the future, we will continue to extract paragraph and discourse level semantic cues. Especially, more high-level semantic information should be taken into consideration. Therefore, we are now trying to develop a rhetorical structure theory (RST) [10]-based hierarchical prosodic model.

Topic 00	country	child	means	books	city	governmer	daily	know	idea	truth
Topic 01	such	may	silence	almost	real	sense	seemed	memory	pain	come
Topic 02	may	most	nor	neither	whole	history	beautiful	latter	certain	nature
Topic 03	sat	hands	came	sid	mary	aunt	face	boy	turned	arms
Topic 04	miles	mississipp	times	between	island	mile	mouth	wide	places	boat
Topic 05	pretty	room	soon	sight	stand	major	town	dark	george	cut
Topic 06	without	name	nothing	come	face	doubt	saying	only	happy	idea
Topic 07	turned	glacier	even	hair	stone	rope	outside	distance	close	guide
Topic 08	hot	cold	dinner	drink	such	tom's	eat	themselves	came	bread
Topic 09	going	company	help	talk	during	knew	return	comfortabl	stay	pretty
Topic 10	came	moved	weeks	became	dropped	procession	hands	land	short	forward
Topic 11	church	big	come	stopped	deck	door	boys	children	town	bell
Topic 12	minutes	five	fifteen	six	miles	boat	thirty	twenty	o'clock	watch
Topic 13	know	read	book	joe	injun	voice	boy	call	because	letter
Topic 14	town	rise	village	looking	bright	spot	miserable	napoleon	came	huge
Topic 15	name	does	should	even	myself	therefore	try	merely	courier	most
Topic 16	chair	breath	floor	boat	caught	thunder	instant	rain	lightning	bed
Topic 17	because	come	anything	only	nothing	seemed	answer	should	else	knew
Topic 18	should	most	far	bad	man's	fine	result	such	between	please
Topic 19	boys	sleep	voice	ladies	talking	together	girls	cried	becky	air
Topic 20	black	blanc	mont	red	remained	smoke	neck	soon	ascent	filled
Topic 21	dollars	year	ago	only	fifty	worth	money	nothing	five	month
Topic 22	sir	only	corps	present	war	sunday	times	seems	fight	number
Topic 23	among	interest	pilots	fire	faces	rose	subject	therefore	mine	even
Topic 24	christian	only	science	isn't	mrs	human	jews	law	perfect	family
Topic 25	know	it's	that's	i'll	can't	come	didn't	i'm	want	won't
Topic 26	far	ground	such	green	trees	beautiful	near	mountains	lake	forest
Topic 27	without	may	however	pilot	alone	boat	god	afraid	full	going
Topic 28	snow	against	wall	point	between	seemed	precipice	summit	passed	valley
Topic 29	sie	ich	die	und	nicht	ist	wirthin	mir	das	herr

Table 3: The topic-term table automatically generated by LDA_based topic modeling method.

Acknowledgements

This work was partially supported by the National Science Council, Taiwan, under the projects with contract 100-2221-E-027-006 and 101-2221-E-027-129.

References

[1]. HMM-based Speech Synthesis System, http://hts.sp.nitech.ac.jp/, Aug. 2012

[2]. Blizzard Challenge, http://www.synsig.org/index.php/Blizzard_Challenge, Aug. 2012

[3]. Stanford Topic Modeling Toolbox, http://nlp.stanford.edu/software/tmt/, Aug. 2012

[4]. The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml, Aug. 2012

[5]. SoX - Sound eXchange, http://sox.sourceforge.net/

[6]. Satoshi IMAI,Cepstral analysis synthesis on the mel frequency scale, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83, 1983.

[7]. D Talkin, A Robust Algorithm for Pitch Tracking (RAPT), Chapter 15, Speech Coding and Synthesis, Elsevier, 1995.

[8]. Y. Itaya, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura. Deterministic annealing EM algorithm in acoustic modeling for speaker and speech recognition. IEICE Trans. Inf. & Syst., E88-D(3):425–431, 2005.

[9]. Y.-J.Wu, R.-H.Wang, 2006. Minimum Generation Error Training for HMM-Based Speech Synthesis, In: Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP), vol. 1, pp. 889–892.

[10]. Mann, William C. and Sandra A. Thompson (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization." Text 8 (3): 243-281.