I²R Text-to-Speech System for Blizzard Challenge 2009

Minghui Dong, Ling Cen, Paul Chan, Dongyan Huang,
Donglai Zhu, Bin Ma, Haizhou Li

Institute for Infocomm Research (I²R), A*STAR, 1 Fusionopolis Way, Singapore 138632

{mhdong, lcen, ychan, huang, dzhu, mabin, hli}@i2r.a-star.edu.sg

Abstract

This paper describes the I²R's submission to the Blizzard Challenge 2009. This is our second participation in Blizzard Challenge. In this paper, we describe the main approaches that we used to build the required voices. We introduce the procedures of database processing, the definitions of the acoustic, prosodic and linguistic parameters, components of cost functions, etc. As our Mandarin system performs well in the evaluation, we will explain more about the Mandarin system. We will also explain how the unit selection method works on very small Mandarin speech database.

Index Terms: speech synthesis, unit selection, cost function, and Mandarin text-to-speech.

1. Introduction

Blizzard Challenge [1] is an evaluation of corpus-based speech synthesis technology on the same databases. It is a good chance to evaluate various speech synthesis methods and is attracting more and more researchers. Blizzard challenge 2009 is the fifth time that this event is organized. This year, the organizer released two databases to the participants. This first one is a 15-hour UK English male speech database provided by CSTR (www.cstr.ed.ac.uk). The second one is a 6-hour Mandarin Chinese female speech database provided by iFLYTEK (www.iflytek.com). Participants can choose to enter the evaluation of one or both languages. For each language, there are hub tasks and spoke tasks. The hub tasks are required from all participants to build synthetic voices, and synthesize a given set of test sentences. The sentences from each synthesizer are then evaluated through extensive listening tests. The optional spoke tasks are: (1) very small datasets, (2) synthesis of speech designed to be heard via a telephone channel, and (3) synthesis of contextually-appropriate speech.

2. Overview of Our Approach

Unit selection approach [2, 3, 4, 5] to speech synthesis has been shown to be one of the best approaches currently. I²R's blizzard 2009 system adopted the unit selection based approach. Both of our English and Mandarin system is based on the same engine.

The first step of building unit selection is the database labelling. In our work, we use automatic forced alignment method using speech recognition technology. We also use some automatic methods to exclude some possible defect units.

Prosody parameters, which include pitch, duration and energy information, are normally used to maintain the naturalness of the synthetic speech. However, spectrally properness of speech unit is also very important for generating good speech. Therefore, in our system, we have a set of acoustic parameters that are designed to cover spectral information in our unit features. The cost function is designed to include these parameters also.

For the Mandarin speech synthesis, instead of using the normally used initial-final definitions of speech unit, we choose to use a smaller phone sized unit. This helps us handle missing syllables easily and makes it possible to generate speech with very small TTS database.

3. Speech Database Processing

In this part, we explain how we process the speech database.

3.1. The Databases

As we have mentioned, the evaluation consists of two databases, they are explained as follows:

English Database: The English speech corpus consists of 15 hours speech in 9,509 utterances, which cover children stories, isolated words, emphasis carrying sentences, news articles, etc[6][7]. The corpus was designed to cover the variants of diphone as much as possible. The corpus comes with transcriptions, which are contained in files of festival utterance format. The RP phone set [8] is used to define the pronunciations of the utterances. There are 50 different phonemes in the corpus.

Mandarin Database: The mandarin speech corpus consists of about 6 hours speech in 6000 utterances. The text transcription of the corpus comes from news corpus. The corpus was designed to cover variants of Chinese pronunciations. The provided information includes: Chinese pinyin pronunciations, prosodic boundary labels, part-of-speech, etc. We defined 43 different phonemes for our task.

3.2. Forced Alignment

In our work, the speech utterances are automatically force aligned with the pronunciations with HTK. Phone-sized speech segment is defined as our basic unit. For the forced alignment, 39 dimensional MFCC feature is used for the training of the phone models. The frame size is 0.025 second and the frame shift is 0.01 second. Three states are defined for each context independent HMM model for each phone. The phone models are first trained with the speech corpus. Unit boundaries are then obtained by forced alignment of speech with its phonetic sequence.

3.3. Unit Filtering

Although the speech corpus is carefully designed and recorded, it is inevitable that some speech units may sound not as good as other units. It is expected that these unit should be excluded in the speech synthesis process. The following measures are taken to remove the possible bad units from synthesis database:

- Use speech recognition: The units are recognized with the HMM models that are trained during forced alignment. If the expected result is not at the top positions, the unit is considered a flawed one and is excluded from the speech synthesis database.
- Use prosody model: When the prosody model is trained, we predict the prosody of the unit from its linguistic features and compare the difference between the predicted values with the actual values. If the difference of two values for a unit is bigger than a threshold, the unit is considered flawed unit.

4. Prosody Model

In this part, we describe how the prosody model of the speech synthesis system is built.

4.1. The Acoustic Parameters

We first define a set of parameters that describe spectral and prosodic features of each HMM state, and boundary frames. The main values that we captured include the statistical values of each individual HMM state, and values of boundary (start and end) frames of the unit. The initial parameters that we used consist of the following:

- Spectral features: MFCC mean for the 3 HMM states, MFCC for boundary frames.
- Pitch features: Mean, maximum, minimum, and range of pitch values and pitch derivative values for 3 HMM states, and boundary frames.
- Duration features: Durations of the 3 states, duration of the unit.
- Energy features: Mean energy of frames in the 3 HMM states, and boundary frames.

Placing all the parameters together, we have a long vector (with a dimension of 308), which contains a lot of redundancy. Therefore, we use principal component analysis approach to reduce the dimension. The dimension reduced vector is considered a compact form of representation of the prosodic and spectral features of the unit. Finally, we have a 40-dimensional vector for both English and Mandarin.

4.2. The Prosodic Parameters

The acoustic parameters are supposed to cover spectral information and prosodic information. However, because there are more spectral information parameters than prosodic parameters in the defined long vector, the prosodic information is actually weakened in the acoustic vector. Therefore, we still need a set of prosodic parameters to emphasize the prosodic properties of speech. The prosodic parameters for each unit consist of the following:

- Pitch mean of the unit
- Duration of the unit
- Energy mean of the unit
- Pitch range of the unit.

4.3. Linguistic Features

The linguistic features are derived from input text. They are used for predicting the acoustic parameters. Due to the difference of language and the information availability, we defined different linguistic features for English and Chinese.

The English corpus comes with the utterance structure for each speech file. We define the features following those that are used in HTS system [9]. We derived the following linguistic features from the utterance files (the number of parameters are given in brackets):

- Context units: phone identities of the previous 2 and next 2 units. (4)
- Syllable information: Stress, accent, length of the previous, current and next syllables. (9)
- Syllable position information: syllable position in word and phrase, stressed syllable position in phrase, accented syllable position in phrase, distance from the stressed syllable, distance from the accented syllable, and name of the vowel in the syllable. (13)
- Word information: length and part-of-speech of the previous word, current word and next word, position of the word in phrase. (12).
- Phrase information: Lengths (in number of words and syllables) of previous phrase, current phrase and next phrase, position of the current phrase in major phrase, boundary tone of the current phase. (8)
- Utterance information: Lengths in number of syllables, words and phrases. (3)

Putting all the features together, we form an input linguistic feature vector of 53 elements for English.

For Mandarin corpus, we defined less linguistic features. The features we used include:

- Context units: phone identities of the previous 2 and next 2 units. (4)
- Tone information: The tones of the current, previous two and next two syllables. (5)
- Phone location in syllable: Number of phones in the syllable, position of the phone counting from left boundary, position of the phone counting from right boundary. (3)
- Word information: length and part-of-speech of the previous word, current word and next word, position of the syllable in word. (8).
- Prosodic phrase information: Lengths of different level prosodic phrases, syllable location in different level prosodic phrases. (12)

Altogether, we have a linguistic feature vector of 32 elements for Mandarin.

4.4. Parameter Prediction

The acoustic parameter prediction process calculates the parameters from the linguistic features. The prediction can be represented with the following formula:

$$y_i = F_i(X) \tag{1}$$

where y_i is the i-th parameter for the unit, X is the linguistic feature vector for the unit.

In our system, the linguistic features are the predictors and the acoustic parameters and prosodic parameters are the responses. We build our models using CART [10] approach. Each individual parameter is predicted separately with a CART tree.

5. Unit Selection

Unit selection method is used in all the voices that we have built. In this part, we describe how we define the cost function.

The unit selection process is based on the cost function that consists of two parts (1) a target cost to measure the difference between the target unit and the candidate unit. (2) a join cost to measure the acoustic smoothness between the concatenated units.

Our target cost further consists of three parts (1) the cost of acoustic parameters, (2) the cost of prosodic parameters, and (3) the cost of context linguistics features. The target cost c_t is defined as the following:

$$c_t = w_{ta}c_{ta} + w_{tp}c_{tp} + w_{tl}c_{tl} \quad (2)$$

where, c_{ta} , c_{tp} and c_{tl} are cost of acoustic parameters, cost of prosodic parameters and cost of linguistic features, w_{ta} , w_{tp} and w_{tl} are the weights respectively.

The reason why we use three cost components here is that each of them alone is not sufficient to describe the target cost. The cost of linguistic feature is to ensure the general spectral and prosodic correctness of the candidate unit. However, due to the variations of speech, using this cost alone may easily leads to extreme cases (abnormal spectrum and prosody). The use of cost for acoustic parameter can avoid the selection of the extreme cases, because the statistical models favor the average values. The use of prosodic cost is to highlight the importance of prosody features.

The cost of acoustic parameters c_{ta} is defined as follows:

$$c_{ta} = \sum_{i=1}^{n_a} ((u_i - v_i)/s_i)^2 \quad (3)$$

where n_a is the dimension of the acoustic feature vector, u_i and v_i are predicted parameter vectors for target unit and the actual parameter vector for candidate unit, s_i is the standard deviation of the i-th parameter.

The cost of prosodic parameters is defined in a similar way to the cost of acoustic parameters. The difference is that weights are added to the cost. The cost is defined as follows:

$$c_{tp} = \sum_{i=1}^{n_p} w_i((p_i - q_i)/t_i)^2 \quad (4)$$

where n_p is the dimension of the prosodic feature vector, p_i and q_i are predicted parameter vectors for target unit and the actual parameter vector for candidate unit, t_i is the standard deviation of the i-th parameter, w_i is the weight of the i-th parameter.

The cost of context linguistic features c_{tl} is defined according to the difference between the features of the target unit and those of the candidate units. When the feature is different, a cost value is given. The total cost is the sum of all the costs for each individual features. In this function, we give higher cost value to the mismatch of important factors (e.g. the identities of immediate previous unit and immediate next unit, the accent of the unit, the stress of the unit, etc).

The join cost, c_j is defined as the squared value of the Euclidean distance between the vector of the end frame in the previous unit E_{i-1} and the vector of the start frame in the current unit S_i as

$$c_j = (E_{i-1} - S_i)(E_{i-1} - S_i)^T \quad (5)$$

The total cost *c* is calculated with the following function.

$$c = w_t \sum_{i=0}^{n} c_t(i) + w_j \sum_{i=1}^{n} c_j(i) \quad (6)$$

where *n* is number of units in the sequence, $c_t(i)$ is the target cost of unit *i*, $c_j(i)$ is the join cost between unit *i-1* and unit *i*, w_t and w_j are weights for target cost and join cost respectively.

The best unit sequence is determined by searching for a best path among the candidate unit lattice to minimize the total cost of the selected sequence. Viterbi algorithm is used to find the best sequence. The weights in the cost function are manually tuned.

6. Building Voices

In this year's evaluation, we have built the following voices: EH1 (full data set), EH2 (arctic data set), ES2 (telephony channel), MH (full data set), MH1 (small data set: 100 utterances), and MH2 (telephony channel). All the voices are built with unit selection methods. It is worth to mention that we have managed to use unit selection for 100 utterances for Mandarin data in task MS1. For telephony voice, we have added an extra post-processing step trying to make it suitable for telephony channel.

6.1. Mandarin Voices

Mandarin is a syllable based language, in which each Chinese character is pronounced as a mono-syllable. There are about 408 base syllables in Mandarin. Each base syllable can be decomposed into an Initial-Final structure similar to the Consonant-Vowel relations in other languages. Each base syllable consists of either an Initial followed by a Final or a single Final. The Initial is the initial consonant part of a syllable and the Final is the vowel part including an optional medial or a nasal ending. In Mandarin Chinese, there are 22 initials (including a null-initial) and 38 finals [11].

Table 1. ***Initial Finals of Mandarin***

22 Initials	b c ch d f g h j k l m n p q r s sh t x z zh null-initial
38 Finals	a ai an ang ao e ei en eng er i ia ian iang iao ie in ing iong iu iz izh ong ou u ua uai uan uang ueng ui un uo v van ve vn

In our system, we further divided the finals into 1-4 phonemes, similar to the phone set used for English speech recognition. Therefore, we defined 43 phones as shown in Table 2. The advantage of using the smaller unit is that we are able to handle missing syllables easily.

Table 2. ***Mandarin phone set***

18 vowels	a aa ah e ea ee een eeng eh er i iz izh o oh oo u v
25 consonants	b c ch d f g h j k l m n ng p q r s sh t vh wh x yh z zh

Because we used a smaller unit size, there are more unit candidates for small data collection. In task MS1, we calculated the number of units in the first 100 utterances as in

Table 3. From the table, we can see that, except for the phone 'oh', all the phones have at least 20 occurrences. As Mandarin is a tonal language, we also examined the tonal vowels, and noticed that most of the frequently used vowels appear in the data set. Totally, there are 10128 units in the small data set. Therefore, it is possible to use the data set as a unit selection database.

Table 3. ***Number of unit s in 100 utterances (For Mandarin voice MS1)***

i	1313	x	180	ch	105
u	1127	v	170	f	103
n	747	een	165	s	80
a	587	izh	164	k	72
ng	575	h	164	r	71
e	435	oo	162	iz	62
d	400	ee	156	vh	58
aa	359	b	154	c	49
ea	305	o	137	p	42
sh	273	t	137	eh	32
j	259	m	122	er	26
l	229	wh	121	ah	20
zh	222	q	117	oh	1
yh	216	z	112		
g	190	eeng	109		

6.2. Telephony Channel Voices

First described by Etienne Lombard in 1911, the Lombard effect is a phenomenon in which speakers alter their vocal production in noisy environments. Measurable differences have been found in vowel duration and intensity in previous research examining the acoustic differences between Lombard speech and normal speech [12, 13]. Prompted by the influence of the Lombard effect on the speakers, we modify the speech prosody, in order to improve the intelligibility of synthetic speech in poor channel conditions.

We have also observed that the unvoiced part of speech can be degraded largely when the speech samples are transmitted via the telephone channel. Therefore, it is necessary to increase its amplitude, in order to preserve the intelligibility of unvoiced parts. As our TTS system is a unit selection based system, we know the exact unit composition of each segment when generating the speech samples. This makes it easy to identify the unvoiced unit without using an unvoiced detection program.

In our method, prosody modification involves increasing the intensity of unvoiced speech segments and lengthening the duration of speech without affecting its naturalness. The trade-off between speech naturalness and comprehensibility is considered when we choose the modification magnitude for increase of intensity and lengthening of duration. We used STRAIGHT synthesizer to extend the duration to 1.2 times of the original speech and amplitude of unvoiced part to 1.5 times of the original value.

7. Results and Discussions

The organizer of Blizzard Challenge 2009 has conducted listening evaluation and released results. This helps us to better understand the performance of the method we used in the system.

This is the second time that we participant in the evaluation. Compared to the evaluation results of last year, we have noticed that the median of naturalness MOS for Mandarin voice (MH) for all listener category improved from 3 to 4. However, the median of naturalness for English voice (EH1) decreased from 3 to 2. The major difference between this year and last year is that this year's system has included more items in the cost function. Therefore, there are a lot of weights need to be tuned. This may make it difficult to achieve the system's optimal performance when tuning the weights. The median of similarity MOS scores of the English and Mandarin system remain the same as those of last year (3 for English, 4 for Mandarin). The results of telephony channel voices are however not good as expected in the evaluation. This shows that the idea of post-processing to increase the duration and energy of unvoiced part may not be a good solution and need further investigation. The method may remove some intelligibility problems for some sounds. However, it may introduce more problems for other sounds.

As we have done quite well in Mandarin voices, we now examine the results of Mandarin voices.

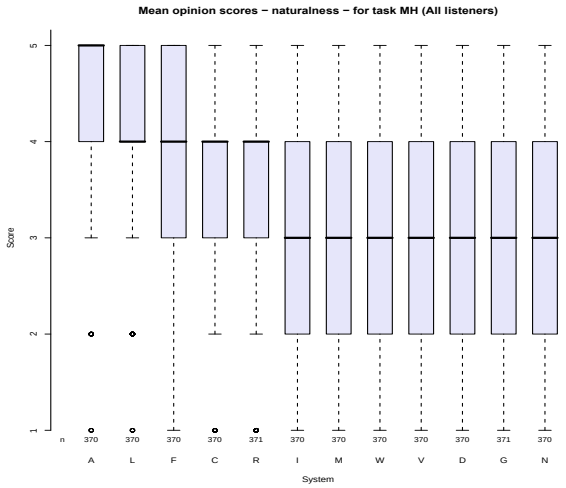

Figure 1. MOS score for Mandarin voice MH (All listeners)

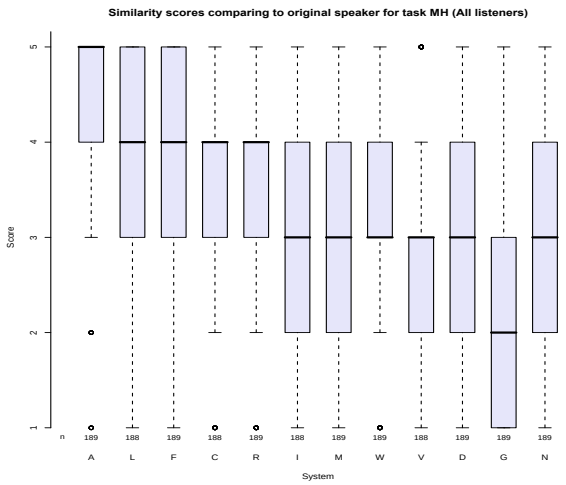

Figure 2. Similarity score for Mandarin voice MH (All listeners)

7.1. Mandarin Voice MH

For Mandarin voice MH, we achieved a mean natural score of 3.5 and a mean similarity score of 3.4. Figure 1 shows the statistics of naturalness score for voice MH from all listeners' feedback. Our system is R in the Figure. From the figure, we can see that our system achieved a median score of 4. This shows that our method is able to achieve high naturalness in synthesizing Mandarin voice. Figure 2 shows the statistics of similarity score for Mandarin voice from all listeners' feedback. From the figure, we can see that our system achieved a median score of 4 for similarity to original speaker. This shows that our method is able to keep the speaker characteristics very successfully.

7.2. Mandarin Voice MS1

When building voice MS1, where there are at most 100 utterances available, we tried the unit selection based synthesis method. To our surprise, the results show that our voice is comparable to other systems.

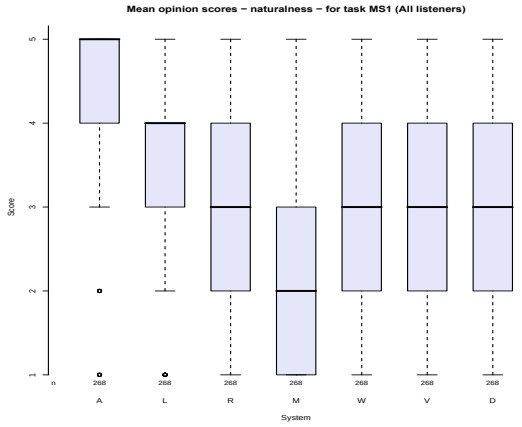

Figure 3. MOS score for Mandarin voice MS1 (All listeners)

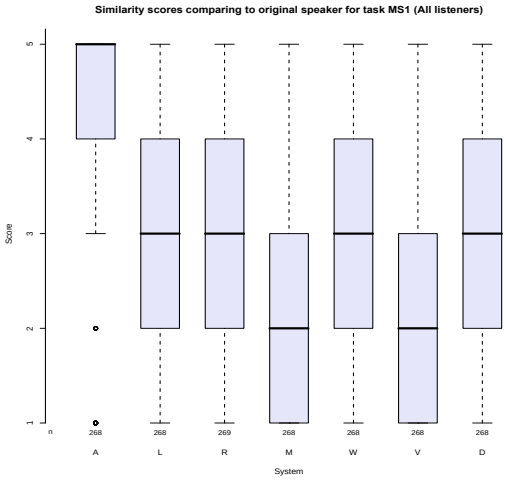

Figure 4. Similarity score for Mandarin voice MS1 (All listeners)

Figure 3 shows the statistics of naturalness score for voice MS1 from all listeners' feedback. Our system is R in the figure. From the figure, we can see that our system achieved a median score of 3. This shows that our method is able to achieve high naturalness in synthesizing Mandarin voice with very small database. Figure 4 shows the statistics of similarity score for Mandarin voice from all listeners' feedback. From the figure, we can see that our system achieved a median score of 3 for similarity to original speaker. This shows that our method is able to keep the speaker characteristics very successfully when using very small database. The success of synthesizing MS1 voice with unit selection method suggests us that, with careful design of speech database, we are able to generate high quality Mandarin speech with very small data set.

8. Conclusion

This paper described our speech synthesis approach for Blizzard Challenge 2009. We used unit selection based approach for all the voices. The evaluation results show that our Mandarin voice is good in both naturalness and similarity. We have also managed to use unit selection for the small database of 100 Mandarin utterances. The evaluation result show the method works well for generating Mandarin speech.

9. References

[1] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King, Statistical Analysis of the Blizzard Challenge 2007 Listening Test Results, Proc. Blizzard Challenge Workshop, 2007, Bonn, Germany.

[2] A. W. Black, P. Taylor, "Automatically Clustering Similar Units for Unit Selection in Speech Synthesis," in Proc. Eurospeech 97, vol 2 pp 601-604, Thodes, Greece.

[3] R. Clark, K. Richmond, V. Strom, S. King, "Multisyn voice for the Blizzard Challenge 2006," Blizzard Workshop 2006.

[4] M. Schroder, A. Hunecke, S. Krstulovic, "OpenMary – Open Source Unit Selction as the Basic for Research on Expressive Synthesis," Blizzard Workshop 2006.

[5] M. Chu, H. Peng, Y. Zhao, Z. Niu and E. Chang, "Microsoft Mulan – a Bilingual TTS System", Proc. of ICASSP 2003, Hong Kong, 2003.

[6] V. Strom, R. Clark, and S. King, "Expressive Prosody for Unit-Selection Speech Synthesis," in Proc. Interspeech, Pittsburgh, 2006.

[7] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, D. Jurafsky, "Modelling Prominence and Emphasis Improves Unit-Selection Synthesis," in Proc. Interspeech, Antwerp, 2007.

[8] S. Fitt, "Documentation and User Guide to UNISYN Lexicon and Post-lexical Rules, Tech. Rep.", Centre for Speech Technology Research, Edinburgh, 2000.

[9] K. Tokuda, H. Zen, A.W. Black, An HMM-based speech synthesis system applied to English, Proc. of 2002 IEEE SSW, Sept. 2002..

[10] L. Breiman, , J. H. Friedman, R. A. Olshen, and C. J. Stone, "Classification and Regression Trees". Monterey, Calif., U.S.A.: Wadsworth, Inc., 1984.

[11] H. W. Hon, et al. Towards large vocabulary Mandarin speech recognition. Proceedings of ICASSP 1994. pp:545-548.

[12] Junqua, Jean-Claude, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," Speech Communication, vol. 20, pp. 13-22, 1996.

[13] W. Van Summers, et al, "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am., vol. 84, no. 3, pp. 917-928, 1988.