

On Unit Analysis for Cantonese Corpus-based TTS

Jun Xu, Thomas Choy, Minghui Dong, Cuntai Guan and Haizhou Li

InfoTalk Technology, Republic of Singapore

{jun.xu, thomas.choy, minghui.dong, cuntai.guan, haizhou.li}@infotalkcorp.com

Abstract

This paper reports a study of unit analysis for concatenative TTS, which usually has an inventory of hundreds of thousand of voice units. It is known that the quality of synthesis units is especially critical to the quality of resulting corpus-based TTS system. This research focuses on the analysis of a Chinese Cantonese unit inventory, which has been built earlier for open vocabulary Chinese Cantonese TTS tasks. The analysis results show that the exercise helps identify the sources of pronunciation deficiency and suggests ways of improvement to address quality issues. After taking remedy measures, subjective tests on improved system are carried out to validate the exercise. The test results are encouraging.

1. Introduction

To build a corpus-based or data driven approach text-to-speech system, one firstly establishes a phonetically rich voice inventory, then employs statistical search technique through the database to find out good voice units in order to compose utterances. In InfoTalk, we have implemented the methodology for mixed and multiple languages. The

study of this paper is based on a Chinese Cantonese system that has been built earlier. The same method was also reported in other commercial products [1][2][3].

Figure 1 illustrates the production process of phonetic mark-up and prosody mark-up of input text into speech. The synthesizer takes the input phonetic and prosody controlling parameters from input text and outputs natural human-sounding speech.

The voice units in the inventory are referred to as *synthesis units*, which originate from a *speech corpus*. Speech corpus is a collection of naturally spoken utterances recorded based on certain balanced scripts, also called *text corpus*, by a voice talent. According to a design specification of the speech synthesizer, these synthesis units are kept in a form of parameters in the storage. There is a notion of a **target cost**, how close a inventory unit is to the desired unit, and a **join cost**, how well two adjacently selected units join together. The synthesizer is designed to optimally minimize both target and join costs, to generate smooth, natural, and fluent human-sounding speech [4].

Assuming that we have a perfect synthesizing, there are still other factors that affect the speech quality [5]. One of them is the database of natural speech, the speech corpus, which forms the foundation of a TTS system. It is more than half way to success if we start with a well-prepared database.

Typically, there are a number of challenges in building a baseline system:

- *text corpus* design
- *speech corpus* recording and labeling
- prosody model training
- text analysis and normalization
- synthesizer development

The reason we call it baseline system is that it is working fine functionally, but it still needs further improvement in terms of voice quality. This routine process allows us to develop new languages rapidly and in a production manner because it is largely automated with the help of speech and linguistic tools and utilities. It is also noted that the rapid prototyping process may produce undesirable defects. As such, a fine-tuning process is needed to exam the synthesis units in a greater details. It is also called unit analysis. Unit analysis will help us identify sources of pronunciation deficiency and suggests ways of improvement.

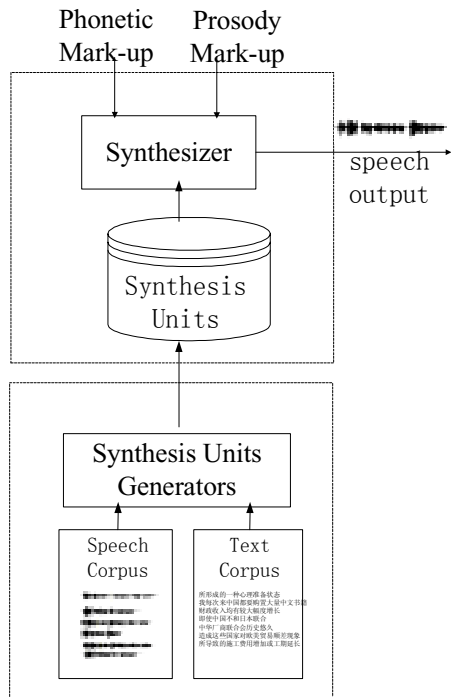


Figure 1. A Corpus-Based Text-To-Speech System

The experiments in this paper focus on the unit analysis of a baseline of Cantonese TTS system. It shows that unit analysis plays an important role in achieving a successful system.

This paper is organized as followings, in Section 2, we introduce the Cantonese TTS system for this study; Section 3 gives the listening test, or subjective test results for the baseline system. Section 4 describes the synthesis unit analysis process. Section 5 shows the improvement of the Cantonese TTS system. Section 6 gives the subjective test results for the improved system. Finally, we conclude in Section 7.

2. The Baseline System

Like other Chinese dialects, spoken Cantonese is seen as a string of monosyllabic sounds. Each Chinese character is pronounced as a monosyllable that carries a specific tone. A character may have multiple pronunciations, and a syllable typically corresponds to a number of different characters. Cantonese is often said to have six citation tones that are characterized by different pitch patterns as shown in Figure 2. The total number of distinct syllables is around 1,800. A Cantonese syllable can be seen to have an *Initial* and a *Final*. In total, there are 19 *Initials* and 53 *Finals* in Cantonese [6].

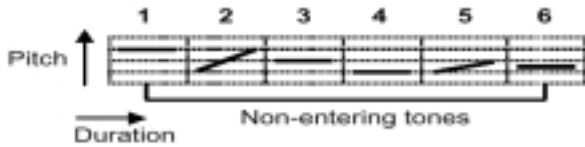


Figure 2. Pitch patterns in Cantonese

In the Cantonese text corpus that we designed, there are all 29,178 sentences containing 290,816 synthesis units, which are recorded in a commercial studio by a professional speaker. The 29,178 utterances form a speech corpus for the TTS system. A synthesis unit is a syllable in size.

3. Subjective Test of The Baseline

Subjective test can be conducted using listening-only methods. In the listening-only method, there are three classes of tests; Absolute Category Rating (ACR) test, Degradation Category Rating (DCR) test, and Comparison Category Rating (CCR) test [7]. In this paper, we used Mean Opinion Score (MOS) test of ACR tests, which asks listeners to rate the speech quality after listening only the synthesis speech without the original speech. In MOS test, listeners are asked to give a rating on a scale of 1 to 5 where 1 represents the utterance having at least one ill-formed unit with inadequate sound or tone, 2 represents the utterance having at least one poor unit with incomplete or partial syllable, 3 represents the utterance having at least one poor unit with abrupt change of speed, volume or pitch, and 5 represents the best quality.

Table 1 shows the distribution of utterances which are rated MOS 3 and below. Those utterances are considered as ill-formed utterances. A total of 45,818 utterances are presented to the listeners and 16.3% of the utterances does not satisfy the listeners and are disqualified.

Table 1: The distribution of ill-formed utterances in baseline system

Domain	Number utterance	Ill-formed utterance	Rate
Finance	2,660	130	4.9%
News	10,000	791	7.91%
HK street names	6,820	558	8.2%
Country and city names	2,039	216	10.6%
Stock names	3,607	526	14.6%
horse racing	1,622	299	18.4%
HK building names	9,038	2,070	22.9%
Person names	6,150	1,631	26.5%
food items	3,882	1,236	31.8%
Total	45,818	7,457	16.3%

4. Synthesis Unit Analysis

An utterance usually contains several synthesis units. One synthesis unit may exist in various utterances. It is reported that 104,368 units are used to synthesize the 45,818 test utterances.

4.1. The Contributing Units

It is found in Figure 3 that only 35.9% of the 290,816 synthesis units in the inventory contributes to the subjective test utterances in the baseline system. This implies that many of the units are not used or are not selected during in the TTS generation process. It suggests that we can reduce the inventory size by a good percentage, which is being carried out in a separate study.

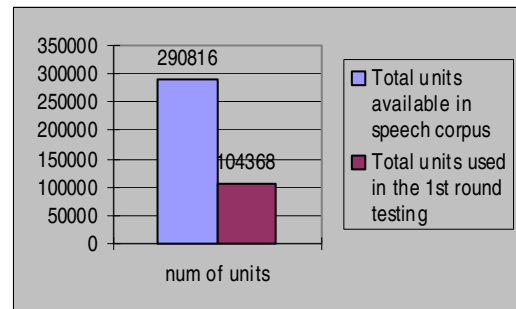


Figure 3. Total units used in the subjective testing

4.2. The Inadequate Units

There are 6,209 suspected ill-formed or inadequate units over the 104,368 contributing units as showed in Table 2. In the following subsections, we exam the distribution of inadequacy of units in different perspectives.

Table 2: Suspected ill-formed units in the baseline system of subjective testing.

Subjective Test	Contributin g Units	Inadequate Units	Rate
Round 1	104,368	6,209	5.95%

4.3. Distribution of Inadequate Units

It is found that only 9% of the total disqualified utterances is related to the issues of lexicon and text-normalization. The trouble units usually have been labelled inadequately phonetically or tonally, which are given MOS score 1.

Over 91% of the total disqualified utterances are related to other issues such as:

- incomplete or partial syllables having MOS score 2
- abrupt change of speed, volume or pitch,
 - syllables heavily affected by the original context in the speech corpus
 - speech too fast or too slow having MOS score 3

4.4. Text Domain of Inadequate Units

It's also found that the synthesis units are good for finance and news domain, but not as good for names and food items. As no spectral or prosodic modification is given to the signal in the baseline system, it is not surprising that even general unit selection synthesizers are still somewhat biased to certain domains.

4.5. Tone of Inadequate Units

Figure 4 shows the distribution of trouble units with tone misinterpretation. It is found that most of the problems caused by tone1, tone3 and tone6 syllables. Cantonese tone1, tone3, tone4 and tone6 have level tones and are quite similar, which easily results in confusion. Tone2 and tone5 have rising tones and are quite distinct from the others. It is observed that tone2 and tone5 syllables have fewer problems. The tone patterns can also be found in Figure 2.

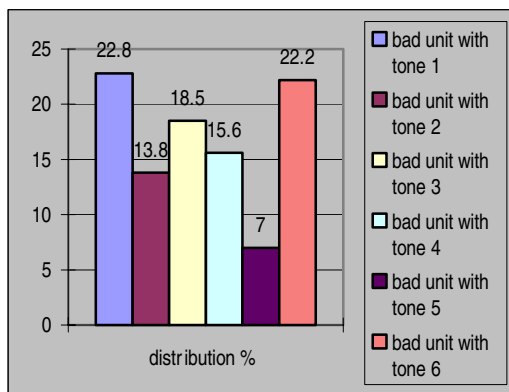


Figure 4. Distribution of trouble units with tone misinterpretation

4.6. Phonetic context of Inadequate Units

It is observed that certain syllable units prone to error in the TTS system. In this section, we summarize some statistics of trouble units concerning phonetic context. It is also found that, although special care is given to the phonetic contextual balance when designing the text corpus, there are still unseen phonetic contexts that occur in the test utterances, which is inevitable.

- The syllables “DAAI6” and “JAU5” account for 2.4% of the total errors.
- The units with Initial “J” account for 14% of total errors.
- The units with Final “AN4” or toneless Final “I” account for 2% and 6% of total errors respectively.
- Units that serve as the leading syllables of an utterance in the speech corpus are prone to error, which account for 18% of the total errors. This may be due to the first syllable of a sample is sometimes having comparatively higher energy/pitch than the others.
- Units that serve as the ending syllables of an utterance are prone to error too, which account for about 4.8% of the total error. The speaker is often out of breath when reaching the last syllables, resulting in comparatively low energy/pitch.
- A Final “I” will have adverse impact on the following sound, resulting in incomplete and uncomfortable units.
- The syllable units in front and after “DIK1” syllable also prone to error.
- The syllable units with Initial “S” also prone to error.

4.7. The Category of Inadequate Units

The above inadequate units can be also categorized:

- Category 1, it sounds bad even in the original sentence in speech corpus, and
- Category 2, it sounds ok in the speech corpus, but not so good in the test.

Inadequate units of Category 1 account for only 6% of the total error.

5. Improvement of Cantonese TTS

The unit analysis also suggests ways for us to improve the voice quality. Sometimes, the issues are interlaced. One inadequate unit may misbehave in different appearing. The quality improvement of a system relies on effective remedy in all the key areas that we identify.

5.1. Improving Units With MOS Score 1

Units with MOS score 1 are related to the issues of lexicon or text-normalization. The missing alternative pronunciations are added to the lexicon address most of the cases.

In addition, user can also define customized rules to deal with non-standard words, such as digit sequence, English letters, roman numbers, abbreviation, currency amounts, acronyms, email address, and so on. These words are not standard in the sense that one cannot find their pinyin properties in the system lexicon [2].

5.2. Improving Units With MOS Score 2

Units with MOS score 2 are related to the issues of incomplete and partial syllables. This is often due to misalignment of the original speech corpus during labelling. It can be alleviated by manually revisit the phonetic alignment and fine-tune it.

5.3. Improving Units With MOS Score 3

Units with MOS score 3 are related to the issues of abrupt change of speed, volume or pitch of the syllables. Improvement is needed within the synthesizer. After

inadequate units are identified, the synthesizer skips all the category 1 units and refrains from using category 2 units.

The tone and phonetics information is used for weighing the joint cost of the synthesizer. For example, a unit coming from the leading syllable of an utterance in the speech corpus should be weighed so as not to appear in the middle of a synthesized utterance.

5.4. Improving Units of Text Domain

It's noted that the synthesis units are not good for names, such as personal names and place names. Listeners expect Cantonese names to be spelt out one-by-one in a different style from reading news.

A name reading mode is added into the synthesizer. In the name mode, the synthesizer would try to find stand-alone units to avoid contextual effect. Similarly practice is also given to other domains that users can identify.

6. Subjective Test of The Improved System

To validate the effectiveness of remedy measures proposed in Section 5, another subjective test is conducted on the improved system. Table 3 shows the distribution of utterances which are rated MOS 3 and below. It is observed that the disqualified utterances are significantly reduced from 16.3%, that of the baseline system, to 6.91%.

Some errors still exist. Most of them are related to units with MOS score 3. It also reflects the quality of the speech corpus in general.

Table 3: The distribution of ill-formed utterances in the improved system

Domain	Number utterance	Ill-formed utterance	Rate
Finance	2,660	66	2.48%
News	10,000	398	3.98%
HK street names	6,820	315	4.63%
Country and city names	2,039	110	5.40%
Stock names	3,607	250	6.93%
Person names	6,150	432	7.02%
horse racing	1,622	146	9.0%
HK building names	9,038	983	10.9%
food items	3,882	465	12.0%
Total	45,818	3,399	6.91%

7. Conclusion

This paper reports a study of unit analysis for concatenative TTS for the improvement of a baseline Cantonese TTS system. It is shown that the unit analysis identifies sources of inadequate units and suggest remedy measures that well improves the TTS quality in subjective tests. The same approach is applicable to development of TTS systems for other languages.

It is noted that general unit selection criteria are not applicable to certain text domain, such as names. Ad-hoc modes are suggested to address the issue.

It is also found that there is redundancy of in the voice inventory, which suggests a reduction of the inventory size is possible.

8. References

- [1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System", in Jont Meeting of ASA, EAA, AND DAGA, Berlin, Germany, 15-19, 1999.
- [2] Jun Xu and Haizhou Li, "InfoTalk Speaker 2.0, The state of the art TTS system", Technical Report HK/SG-2002, InfoTalk Research and Development Center, Singapore.
- [3] P. Rutten, G. Coorman, J. Fackrell & B. Van Coile, "Corpus based speech synthesis in the Lernout & Hauspi RealSpeak TTS system," Proc. IEE symposium on State-of-the-Art in Speech Synthesis, Savoy Place, London, pp.16/1-16/7, 2000.
- [4] A. Hunt and Alan W Black, "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database", in ICASSP-96, Atlanta, Georgia, 1996, vol. 1, pp. 373-376.
- [5] Alan W Black, "Perfect synthesis for all the people all of the time", Proceeding of IEEE 2002 Workshop on Speech Synthesis.
- [6] Tan Lee, Greg Kochanski, Chilin Shih and Yujia Li, "Modeling tones in continuous Cantonese Speech", Proceeding of ICSLP 2002.
- [7] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," International Telecommunication Union, 1996.