Subjective intelligibility testing of Chinese speech

Ian McLoughlin

Abstract

This paper presents a complete methodology and rationale for the subjective intelligibility testing of Chinese speech. It replaces the combination of several previously published Chinese intelligibility tests which have been in use for almost a decade, with a single composite test procedure constructed from a foundation of subjective trials and auditory evidence. Since publication of the first elements of Chinese intelligibility test, several factors have come to light which prompted this overhaul. Firstly, international testing has highlighted words used in the original test that are unsuitable for speakers of particular regional dialects. Secondly, recent evidence indicates that the assumptions of tonal confusion made during the definition of the original tonal intelligibility tests are not borne out by subjective evidence. Finally, words published in the original test disadvantaged speakers from Mainland China due to the use of full-form Chinese characters rather than the more ubiquitous simplified form characters. This paper presents experimental evidence of tone confusion in Chinese speech, and uses this data to create a replacement tone test. Word choice has been adjusted to find more neutral alternatives for particular regional dialect speakers. The basic speech and tone extension tests are now presented with simplified form characters to ensure accessibility by the greatest number of test subjects. Finally, this paper includes a description of the full intelligibility test.

Index Terms

Mandarin, Chinese, intelligibility, DRT, CDRT, tone
Subjective intelligibility testing of Chinese speech

I. INTRODUCTION

Speech processing technologies underpin an important aspect of daily life for millions of people worldwide, particularly through cellular communications, internet telephony and video conferencing. Even the humble POTS telephone network is increasingly relying on digital coding. Plans by British Telecom to turn off the analogue telephone network in the next few years are likely to be replicated elsewhere and lead to all-digital telecommunications. Meanwhile, despite near-saturation in several Western countries, adoption of cellular telephones continues unabated worldwide.

Coincident with this backdrop is the rise of China as an economic superpower, and the resulting adoption of personal telecommunications technology among the 1.3 billion strong population. Already, rates of cellular telephony ownership in China are significant, and along with internet use, continue to grow rapidly. In all likelihood, there will soon come a time when the majority of worldwide speech processing is operating on Mandarin speech. From a commercial viewpoint, Mandarin speech communication is likely to constitute the world's largest telecommunications market.

At present, there are many standards for assessing the quality and intelligibility of speech conveyed by a network or processed by a speech handling system, with several being enshrined in ITU or ANSI standards, most particularly ANSI standard S3.2 [1].

In general, methods of speech testing are classed as being either subjective or objective: requiring a group of human listeners, or conducted by automated systems, respectively. Tests determine either quality (how nice the speech sounds) or intelligibility (understanding of the items spoken): different communications systems tend to aim towards optimisation of either one or the other, but not both. For example, emergency service communications requires high levels of intelligibility whereas broadcast audio services tend to sell better if quality is high.

Reasonable automated quality measures such as PAMS, PESQ and PSQM already exist, but intelligibility testing is particularly challenging to assess through automated means, and thus human listening tests are often required. The speech transmission index (STI) and related automated methods [2] are promising, but have not yet been proven capable of completely replacing traditional subjective testing. One of the more common subjective methods, mentioned in ANSI S3.2, is the diagnostic rhyme test (DRT), an A/B forced comparison test based on word pairs differing by a single attribute (see section III) [1]. This method, originally published in 1983 by Voiers [3], has been used by the author for many years, and generally found to demonstrate good repeatability and accuracy. However the test is only defined for English speech [1] (in fact it is a subset of this: English speech with an American accent). Quality tests, by contrast, tend to be language independent.

In 1999, the methodology described by Voiers in the original definition of the DRT was repeated in a Chinese context, to generate a Chinese diagnostic rhyme test (CDRT) [4]. This has since been used in the evaluation of the GSM 06.10 speech coder [5] [6] and ITU G.728 [7] for conveyance of Mandarin speech. Results for both codecs indicate slightly poorer performance compared with their handling of English speech. Such evidence has been used to propose the development of speech codecs that are specifically optimised for handling Mandarin, in contrast to the majority of codecs in use today, which are predominantly Euro-centric: originally developed using English or French speech, and optimised for those languages.

Despite the development of a DRT for Mandarin speech, there are major differences between Chinese and English (discussed in section II) which were not well catered for in the methodology of the test. These prompted research which led, in 2001, to a proposed extension of the CDRT to cater for tone discrimination effects [8]. The CDRT-tone was then used to evaluate the tonal transparency of several speech codecs including GSM 06.10, and used to propose improvements to the pitch handling mechanism in the GSM RPE coder specifically for Mandarin speech [9].

In the original definition of the CDRT-tone test, the precise choice of tone combinations used in the test was based upon a visual examination of the average tone-track plotted for several hundred spoken Chinese characters: those tracks exhibiting trajectory similarities were considered to be more easily confused, and therefore included in the test. In fact, new data now reveals that the human perception of tone does not necessarily follow a physical interpretation of the tone tracks. The original assumptions do not match the experimental evidence, and thus the CDRT-Tone test is not at all optimal. This paper presents new experimental evidence, constructs a replacement test, and evaluates the result.

Experience with the existing Chinese intelligibility tests has highlighted a number of shortfalls, and thus a New Chinese Diagnostic Rhyme Test (NCDRT) is proposed, explained and evaluated in this paper. Section II will discuss the features of the Mandarin Chinese language, and identify areas of difference compared to English. Section III discusses the (English) diagnostic rhyme test procedure, methodology and rationale. Section IV then presents experimental and human factor evidence relating to the previously published Chinese language intelligibility tests, while section V discusses construction and performance of the composite replacement test, comprising NCDRT part I for phoneme intelligibility and NCDRT part II for tonal intelligibility. The test methodology is described in

1 The term ‘Mandarin’ is often used to refer to the majority Chinese dialect, and is used interchangeably with the term ‘Chinese’ in this paper.
section VI whilst section VII concludes the paper.

II. MANDARIN CHINESE SPEECH

WRITTEN Chinese consists of a sequence of unique pictograms called characters. Each character has a particular meaning and pronunciation, and may exist alone, or be aggregated into compound strings to provide a different meaning. In written sentences there is often no typographical separation between neighbouring characters: context and reader experience alone determine when characters are to be interpreted singly, when they are to be grouped, and if so how large is the grouping.

Many characters have several meanings which may be totally unrelated, particularly in the case of the simplified jiantizi form characters used in the People’s Republic of China (PRC) and elsewhere. These are derived from the traditional fanti-zi or full-form, characters in use in Taiwan and Hong Kong, through a method of simplification that reduced the number of pen or brush strokes required for writing them.

Many of the simplifications resulted in totally different complex characters being reduced to identical simplified characters, however the pronunciation between the two character forms (and meanings) remained unchanged. For these characters, and several others, two or more alternative pronunciations exist. The majority, however, support a single pronunciation.

We now consider some of the major differences between English and Chinese speech, which relate to the issue of intelligibility testing.

1) Firstly, all characters are monosyllabic. Thus words that consist of a single character are also monosyllabic. Words or phrases consisting of two characters are bisyllabic: three character words are tri-syllabic and so on. So we can consider the syllable as the basic unit of spoken Chinese, and specify the phonetic properties further [10] [11] [4]. The monosyllabic nature of the language may be an advantage in intelligibility testing, compared to English. For example, the DRT, describes it’s category. Tone 1 has a high-level frequency, tone 2 is mid-rising (meaning it starts at a middle frequency with an upward trajectory), tone 3 is mid-falling-rising and tone 4 is high-falling. Evidently descriptive words such as ‘mid’ and ‘high’ relate to perception rather than a particular physical measurement, but it could be argued that these terms are to be understood in relation to the average pitch frequency of an utterance. In fact, a numeric method of representing tone dating from the 1930s is presented in [14], where the pitch range is divided into five levels, with 1 being the lowest, 3 being mid-range and 5 the highest. In this way, the tones can be described by a sequence of pitch levels as shown in table II.

2) Secondly, each syllable consists of a consonant-vowel-consonant (CVC) structure, although in a few cases the initial consonant is missing (null), and in many cases the final consonant is missing. The final consonant is always nasal (N), being either /n/ or /ng/. Syllables exist that conforming to all combinations of CVN, CV, V and VN, with a total of about 415 permutations allowed [4] (although [11] mentions only 408). Initial consonant and vowel (including the optional final consonant) belong to the allowed set listed in table I. Note that this shows the Chinese Phonetic Alphabet (CPA) transcription normally used for the hanyu pinyin romanisation of Chinese characters [10]. A conversion table between the CPA and the International Phonetic Alphabet (IPA) is given in [4].

Spoken Chinese contains even fewer voiced consonants than English, and these are presented in table I, in bold.

The lack of voiced consonants is an attribute shared by English, but the effects will be more pronounced in Mandarin Chinese. In particular because unvoiced consonants are spoken with lower average power than other phonemes [12].

Relating back to written characters for a moment, given that an estimated 13,000 different characters exist in written Chinese, with only 415 phonetic permutations, it follows that Chinese features a very large number of homophones.

3) The third major difference between Chinese and English is in the use of tonal information. Each spoken Chinese syllable is identified by the combination of phonetic and tonal attributes. Neither alone is generally sufficient to identify a particular word. For example, the same syllable /ma/ can have meanings as diverse as ‘horse’, ‘mother’, ‘numb’, ‘linen’, ‘scold’ or could indicate that the preceding phrase was uttered as a question, with each meaning differentiated by tone.

In Mandarin there are four distinct lexical tones plus a neutral tone often described as lack of tonal contour [13]. Each tone has a particular pitch contour, or track, that defines it’s category. Tone 1 has a high-level frequency; tone 2 is mid-rising (meaning it starts at a middle frequency with an upward trajectory), tone 3 is mid-falling-rising and tone 4 is high-falling. Evidently descriptive words such as ‘mid’ and ‘high’ relate to perception rather than a particular physical measurement, but it could be argued that these terms are to be understood in relation to the average pitch frequency of an utterance. In fact, a numeric method of representing tone dating from the 1930s is presented in [14], where the pitch range is divided into five levels, with 1 being the lowest, 3 being mid-range and 5 the highest. In this way, the tones can be described by a sequence of pitch levels as shown in table II.

Smoothed pitch-frequency contours of the four tones are
drawn in Fig. 1, with the y-axis representing frequency and the x-axis representing time (from data presented in [15]), and the shapes can be compared to the numerical pitch representations of table II. It should be noted that significant deviations from average occur, even for repeated utterances by a single speaker.

Applying tone to the example syllable used previously, it is now possible to better differentiate meanings: thus /ma1/ means ‘mother’, /ma2/ means ‘numb’, /ma3/ means ‘horse’, /ma4/ means ‘scold’ and the neutral /ma/ is the second syllable of the compound bi-syllabic word meaning ‘mother’ (/ma1//ma/). Such an example confirms the degree of importance of tone in Mandarin Chinese, in complete contrast to English which may use tone to convey emotion, but not normally to distinguish word meaning.

The neutral tone is far less important in its tonal presentation, and has been characterised [14] as being half-low when following tone 1, middle-level when following tone 2, half-high when following tone 3 and low when following a tone 4. Note in table II that some tones will change slightly if preceding a neutral, and the neutral will then follow this adjusted tonal endpoint of the preceding tone.

In fact the context-sensitive nature of tonal pronunciation extends to two further cases, called the rules of tone sandhi, affecting tone 3 when spoken in combination with other tones. In continuous speech, tone 3 has the most diverse shape of the Chinese tones [16], in that it is spoken as tone 2 when it is followed by another instance of tone 3, and changes to a low level when followed by an instance of tone 1, tone 2 or tone 4. From table II it can be seen that the final rise is also not present when tone 3 is followed by a neutral tone, and thus the only time when tone 3 is uncorrupted is when it is spoken either in isolation or in an utterance-final position [14].

Tone 3 is thus different when presented in isolation compared to when it is juxtaposed in speech [17], and this fact is particularly significant since the DRT test, as we shall see in the next section, uses single words to determine the intelligibility of speech, rather than continuous sentences.

III. DIAGNOSTIC RHYME TEST

A NSI S3.2 is intended “for use in measuring the intelligibility of English speech”, describing a diagnostic rhyme test in which a number of trained speakers read a list of words which are presented to listeners both directly and through an audio handling system under test. The difference between the two provides information regarding any degradation caused by that system.

The DRT test material comprises a word list of 96 rhyming pairs such as /bat//hat/. Each of these 192 words is spoken in random order. One set of words is used as a reference, while the other set is passed through the system under test. A panel of listeners, who have been trained to a plateau, are then presented with the 384 words (192 reference and 192 processed) in random order, and for each word are required to indicate what they heard. They are given an A/B forced choice consisting of the spoken word and its rhyming pair. In DRT, the words of each pair differ only in their initial consonant (a more lengthy sister test, the modified rhyme test, or MRT, uses a similar method for a six-alternative forced choice, with half the word set differing by final and half differing by word initial).

In the DRT, the proportion of correctly identified words is tallied for both the reference and the processed set. In general terms, a test system that degrades the audio minimally will yield a similar percentage of correct responses to that exhibited for the reference words. A system that significantly degrades the audio will display a far lower percentage of correct responses than are exhibited for the reference words. DRT word pairs are divided into six classes based upon the phonetic attribute that differs between the alternative initial consonants - with the per-class result breakdown providing greater insight into the fidelity of several different parts of speech handled by the system under test.

In typical testing scenarios, subjects are trained to a plateau in advance by prepending a selection of the rhyming pairs to the start of the test, presenting these for testing, but not including the answers in the analysis results. A set of calibration trials is usually performed prior to the actual test to determine the optimum speech amplitude to be presented to listeners.

The CDRT is similar to the DRT, apart from the use of recorded Chinese words, and the presentation of Chinese characters (or hanyu pinyin, the standard phonetic transcription used in PRC and elsewhere) for the scoring. Obviously both the speakers and the listeners should be familiar with the language under test. The choice of phonetic attributes used in the original CDRT test are given in [4].

The CDRT-Tone extension [8] uses the DRT methodology to evaluate tone discrimination: but instead of rhyming word pairs differing by initial consonant, the words differ in their tone. Thus {/ma3/, /ma4/} would constitute such a pair. In total, four groups of ten tone pairs were included in the original published tests, chosen from an examination of the pitch tracks of various characters as stated here [8]:

“According to the similarity of pitch height and contour among the four Chinese tones, four categories {tone 1 - tone 2}, {tone 1 - tone 3}, {tone 2 - tone 3} and {tone 3 - tone 4} are chosen to form the basis of the tone intelligibility test.”

In the underlying work [18] used to derive the CDRT-Tone, examination of tone tracks, similar to those plotted in Fig. 1, were used to determine visual similarities between plots of spoken tone tracks using several attributes. These are presented in table III, showing that the first three chosen test categories of {tone 1 - tone 2}, {tone 1 - tone 3}, {tone 2 - tone 3} display a marked similarity and were thus reasonably assumed to be susceptible to confusion. The {tone 3 - tone 4} category however should probably have been replaced (according to the table) by {tone 1 - tone 4}, and in fact the empirical evidence of section IV bears out such a conclusion. The use of {tone 3 - tone 4} was due to initial tone contour similarity [8], in that tone 3 and tone 4 are the only tones to exhibit a tone transition from high to low, and were thus assumed to be more prone to confusion.

While the rationale for the test choice was logical, evidence indicates that humans do not confuse tones strictly according to the physical criteria used. In fact several studies indicate that the location of the turning point in the pitch contour is
TABLE III

VISUAL SIMILARITY OF CHINESE TONE PITCH CONTOURS

<table>
<thead>
<tr>
<th>Heard as:</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1 spoken</td>
<td>h</td>
<td>c</td>
<td>c</td>
<td>h</td>
</tr>
<tr>
<td>Tone 2 spoken</td>
<td>c</td>
<td>-</td>
<td>c</td>
<td>h</td>
</tr>
<tr>
<td>Tone 3 spoken</td>
<td>c</td>
<td>c</td>
<td>h</td>
<td>-</td>
</tr>
<tr>
<td>Tone 4 spoken</td>
<td>h</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

h: similarity in pitch height  
c: similarity in pitch contour  
-: no significant similarity

a particularly important feature in the discrimination of tone [19] as is the relative pitch between onset and turning point [20]. Duration also plays a part [21], and none of these features had been considered in the CDRT-tone design. While relative pitch discrimination in humans can be good, it seems that rising tones are in general more difficult to discriminate than falling tones [14]. This last fact was substantiated by Klatt [22] who reported an unpublished experiment by Victor Zue which progressively flattened the tonal contours of a synthesised word while measuring subjective tone confusion, until, with a 97.5% reduction in tone gradient, the two rising tones {tone 2 - tone 3}, became almost indistinguishable, whereas the {tone 1 - tone 4} difference remained identifiable.

Several other studies have investigated the confusion of Chinese tone under less extreme levels of distortion. In particular, the 1976 study by Howie [15] reported extensive results on 34 representative syllables of nine basic types. Although in many cases the confusion results given were clearly within the margins of error, there are three results worth highlighting from the 12-listener tests on the presentation of a single-speaker synthetic word /bao/.

Firstly that {tone 1 - tone 4} confusion, missing in the original CDRT-Tone test, is seen in the Howie data [15] albeit at a low level, and thus probably should be included in any test of tonal intelligibility. Secondly that, {tone 1 - tone 3} and {tone 3 - tone 4} confusion, comprising one third of the original CDRT-Tone test, was almost non existent when the words were not vocoder processed (distorted).

Finally, and most importantly, is the issue of {tone 2 - tone 3} confusion. In a test repeated with and without carrier sentence (Zhei4 ge { } zi, shi Lao3 Li3 xie3 de) [15], the confusion rate of tone 3 as tone 2 increased by a multiple of 13 (by contrast the confusion of tone 2 as tone 3 barely changed, but was eclipsed by the confusion of tone 2 as tone 1). In fact this is due to the issue of tone sandhi (see section II) – when the tone of the word under test was adjusted at will to perform the test, the carrier sentence remained unchanged, thus resulting in a sentence that was “never heard in real speech” [15].

Shen and Lin [23], in a precursor to their 1993 paper [19], considered tones 2 and 3 which they believed more confusable due to their similar concave shapes. When attempting to recreate earlier reported test results which showed tone 3 mistaken as tone 2 to be more prevalent than tone 3 mistaken as tone 2, they obtained results contrary to expectations. This may be a classic issue of tone sandhi, but even so, their significant data points derived from just three specific syllables which were often misidentified. Their conclusion being that since some words are inherently more confusible than others, they had simply chosen a different word set to the earlier experimenters and thus obtained different results.

Clearly, the confusion expectations used in the original CDRT-Tone test design do not match the published evidence, and furthermore the published evidence itself differs markedly depending upon the type of distortion present, effects of tone sandhi evident with the use or non-use of a carrier sentence, and the particular words chosen for the smaller tests.

A new tone test will therefore constructed be in section V. This will now test all combinations of tone confusion, and will do so in a proportion based on subjective evidence. The subjective evidence, obtained in section IV, will be based upon single DRT-style words presented individually and thus free from the effects of tone sandhi. The test corpus will comprise a large number of words, chosen across the different phonetic groups, and will be corrupted by additive auditory noise, rather than vocoder-style tonal manipulations.

IV. NEW EVIDENCE

In order to evaluate tone confusion, a set of 260 Chinese words were chosen across the four non-neutral tones, and encompassing various word structures including consonants,
vowels, nasals and glides \{CV, CVN, VN, CGV, CGVN\} as shown in table IV.

The criteria for test word choice was to include commonly used characters, and in particular homophones spanning each of the four non-neutral tones (in this instance ‘word’ refers to the hanyu pinyin romanization excluding tone). Characters were presented so that otherwise identical words were presented several times but with different tone, such as /mi1/, /mi2/, /mi3/ and /mi4/. For some words, common characters could not be found to span all four tones (as judged by a panel of native Chinese speakers), and in such instances, only the two or three instances of the word in common use were included in the test. To compensate, some words were included more than four times where common examples of the same word could be found. For example, two instances of /ba1/ and one each of /ba2/, /ba3/ and /ba4/ - each representing a different Chinese character. On average, slightly more than four presentations per word were achieved. With the intention of measuring the degree of perceived tonal confusion between each pair of tone possibilities, the test thus spanned the six possible tone pairs:

- \{tone 1 - tone 2\}
- \{tone 1 - tone 3\}
- \{tone 1 - tone 4\}
- \{tone 2 - tone 3\}
- \{tone 2 - tone 4\}
- \{tone 3 - tone 4\}

The test material was prepended by a set of training words and spoken by an announcer chosen on the basis of demonstrated tonal clarity, and recorded to computer. The female announcer was from mid-China, and spoke standard Mandarin. White Gaussian noise was added to the speech in various degrees of signal-to-noise ratio (SNR), and an initial calibration test run performed with a group of five Chinese speaking listeners to determine the optimum level of SNR required to provide approximately 10% to 20% incorrect recognition rate, and to prove the experimental methodology, consisting of a pencil-and-paper test conducted individually through headphones. For the main test, symphonic music (Liang Shan Bohe Zhu Ying Tai [24]), was used to corrupt the speech at a segmental signal-to-noise level equivalent to that determined through the calibration tests. More specifically, the power of each individual uttered word was measured, compared to the power of noise in the region to which it was to be added, and then amplified to obtain an SNR as specified by the initial calibration tests. In this way, the SNR of each test word was individually normalized against the corrupting noise before presentation to listeners. White noise was not used in order to reduce effects of listener fatigue over the lengthy test. A second five-person calibration test using the same material was performed to ensure that the use of non-white noise did not affect tone discrimination in any significant or consistent manner.

For the main test, one hundred Chinese speaking volunteer listeners, approximately 21 years old, were each presented with the words, spoken in random order, and given a forced choice answer sheet. Listeners were not paid for the tests, reported normal hearing, and were familiarised with the procedure prior to test commencement. The condensed results from this test, in terms of confusion between spoken and heard tone, are shown in Table V and plotted as a bar chart in fig. 2.

It can be seen that tone 3 is the most easily recognised tone (as was predicted in [8] due to it’s large shape difference compared to other tones), correct in approximately 2% more cases than any other tone. However if we examine the most confused sets of tones, these do not match up with the other predictions made in [8].

To examine further, the first column in table VI shows the overall percentage confusion of the various tones in the test, and is presented in decreasing order of most confused tone pairs. The four tone pairs included in the original CDRT-Tone, namely \{tone 1 - tone 2\}, \{tone 1 - tone 3\}, \{tone 2 - tone 3\} and \{tone 3 - tone 4\}, are given in bold.

It is quite clear that three of the CDRT-Tone tested categories were actually the least confused tone pairs and furthermore, the pair that is most likely to contribute to misunderstanding is omitted entirely from the original CDRT-Test.

### V. TEST CONSTRUCTION

The basic CDRT test has now been well used worldwide for many years, and whilst indications have generally been positive regarding repeatability and word choice, the original test was presented and published in fanti-zi (full-form characters) [4] which are not generally taught in the PRC, Singapore or Malaysia. This paper has thus converted the test material to the far more widely used jianti-zi simplified form. The pronunciation, and hence hanyu pinyin romanization are maintained, however since the characters are now accessible to a far wider readership, they allow for a greater choice of test subjects.

In experimental testing designed to deduce a relationship between speech quality and Chinese speech intelligibility [25] [26], the existing CDRT was performed on many overseas Chinese listeners located in Christchurch, New Zealand. Among the test group were 8 native speakers from Taiwan. Test results indicated that due to different pronunciation, these
TABLE VI

ORDERED LIST OF TONE CONFUSION, SHOWING NUMBER OF PAIRS IN ORIGINAL CDRT-TEST (IN BOLD), AND THE NUMBER PROPOSED FOR THE REPLACEMENT TEST

<table>
<thead>
<tr>
<th>Tone confusion</th>
<th>Percentage</th>
<th>Original</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1 - Tone 4</td>
<td>8.67%</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Tone 1 - Tone 2</td>
<td>7.70%</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Tone 2 - Tone 4</td>
<td>7.39%</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Tone 1 - Tone 3</td>
<td>2.37%</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Tone 2 - Tone 3</td>
<td>2.24%</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Tone 3 - Tone 4</td>
<td>0.58%</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

listeners were unable to clearly distinguish between the {zan4 - zhan4} pair in the sibilated-unsibilated category, and thus the replacement pair of {zui4 - zhui4} has been chosen instead. Small group testing has since confirmed that this pair is more distinguishable to native Taiwanese speakers. No significant difference in CDRT result were noted for Singaporean and Malaysian speakers of Chinese as against a PRC reference. The full list of words used is termed the NCDRT part I, and is given in full in fig. 3.

Table VI, discussed previously, identified the number of each tone pair that is included in the original CDRT-Tone test, and presented the number that should be tested if pairs are to be tested in proportion to their degree of confusion (with the overall test corpus of 40 pairs being maintained).

Using this methodology, a replacement set of words has been determined, which includes tone confusion pairs in proportion to their degree of confusion. This new word set, named the NCDRT part II is given in fig. 4. The characters, all in common use in the PRC, are presented in both jianti-zi simplified form and hanyu pinyin, and are easily recognised in the absence of noise.

Initial indications, using the NCDRT part II, shows that the words scale well with the degree of occluding noise applied,
Fig. 3. Listing of hanyu pinyin and simplified-form Chinese characters forming the six word attribute categories of the NCDRT part I and are more representative of human tone confusion than are the original test words. To examine further, subjective testing was performed using the new NCDRT part II words. This involved five listeners in a headphone-based pencil-and-paper test. Words were occluded by AWGN and presented at two second intervals.

The results are shown in table VII and indicate that the tone 1 - tone 4 category, missing in the original CDRT-tone test, is clearly the cause of a significant degree of confusion. A more detailed analysis of the results indicates that errors were spread widely among the tone pairs. This is shown in fig. 5 which plots a histogram of the number of listeners responding erroneously to the same word pair.

The new test material presented in this section will be explained in the context of the overall NCDRT test in the following section VI.

TABLE VII
NCERT- PART II TONE TEST RESULTS

<table>
<thead>
<tr>
<th>Heard as.</th>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 3</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1 spoken</td>
<td>89.63</td>
<td>3.70</td>
<td>0.00</td>
<td>6.67</td>
</tr>
<tr>
<td>Tone 2 spoken</td>
<td>2.61</td>
<td>90.43</td>
<td>0.00</td>
<td>6.96</td>
</tr>
<tr>
<td>Tone 3 spoken</td>
<td>3.33</td>
<td>3.33</td>
<td>93.33</td>
<td>0.00</td>
</tr>
<tr>
<td>Tone 4 spoken</td>
<td>7.50</td>
<td>2.50</td>
<td>0.00</td>
<td>90.00</td>
</tr>
</tbody>
</table>

VI. NEW CHINESE DIAGNOSTIC RHYME TEST

BASED upon the material presented in the preceding sections, the NCDRT has now been defined. The word lists are shown in figs. 3 and 4 respectively, and this section...
will provide an overview of the test methodology and results analysis. Much of the methodology is deliberately based upon the original ANSI standard for intelligibility testing of English speech [1], since this is known to have characteristics of both repeatability and overall accuracy.

The intention of this section is to provide a self-contained test and methodology for the intelligibility testing of Chinese speech.

A. Test overview

The NCDRT is best applied as a comparative test - highlighting the difference between a recording of reference speech and the same speech degraded by a speech processing or communications system. Unlike in the original ANSI recommendations [1], we will assume that in all cases the speech under test is pre-recorded and pre-processed prior to the listener tests.

The first point to note is that modern communications systems can easily yield a perfect intelligibility score, and this would reveal nothing about any degradation caused by the system. In such cases the reference speech itself must be artificially corrupted with noise prior to testing. In broad terms this moves the operating point to one of slightly lower overall intelligibility, revealing borderline intelligibility features in the speech that may otherwise be hidden. The intention is to ensure that neither the original set of words, nor the degraded audio, yield a response score near to limits of either 100% or 50%. These two extremes correspond respectively to audio fidelity so excellent that no perceptible degradation has occurred, and to such poor fidelity that the result score is indistinguishable from guesswork.

The pre-processing method is shown in fig. 6. Word lists are read out by a number of speakers and recorded. Ideally the number of speakers should match the number of listeners according to Voiers [3], and should include both male and female speakers. In practice however, fewer speakers than listeners are generally employed in the tests. The speaking environment should be quiet and echo-free, and the recorded speech should not be so loud as to clip or so quiet as to be too noisy. A rule of thumb from the recording industry is that when speaking normally the recorded speech level should be around 12dB below full scale.

The set of recorded multi-speaker word lists are then corrupted by noise for the reasons mentioned above. Generally a brief calibration test is needed to determine the level of noise corruption, ideally moving to a level that is sufficient to reduce the intelligibility score a little, down to approximately 80% to 90% correct.

The noise-corrupted word list is considered to be the reference in this test. A copy of the reference is then passed through the system under test (SUT), the speech processing or transmission system that is being evaluated, to yield a processed word list. The reference and processed lists then form the basis of the test material. NCDRT part I consists of 80 word pairs, and thus 160 words in the full word list. Ideally these have been uttered, and recorded by each speaker. Thus for a two speaker system there are 320 basic words. These would then be corrupted by noise to provide the reference, and would form half of the test material. The other half of the test material would be the reference words processed by the SUT. The test material to be presented to listeners thus comprises 640 words in this instance.

If the word list has been processed as an entire recording, it will need to be split into individual words prior to testing, so that these can be randomly presented to listeners. Of course they could have originally been recorded in random order, but it is still necessary to interleave presentation of the reference and processed set in some way.

The actual listening test is conducted as shown in fig. 7. The
test material of noise corrupted words, and processed noise corrupted words, is to be presented to a listener or panel of listeners in random order, constrained to be presented an equal integer number of times each (for example two presentations of each word - effectively doubling the test duration).

Each listener is given a two-alternative forced choice when being presented with each replayed word, and thus a string of guesses would average out at a 50% correct score. For this reason, the final results are corrected for guesswork.

In terms of scoring, the main result is the difference between the percentage of correct responses for words which have been processed by the SUT and words which have not. Each of the six DRT or NCDRT categories also yields separate scores, instructive in determining what word features are most affected by the SUT. For example in [5] a large score reduction in the CDRT {sibilated-unsibilated} group reveals that the SUT, the GSM 06.10 codec, does not preserve sibilation features as well as it preserves other tested features of speech. Improvements to the codec could then be directed at this aspect of the system.

Further detail on the test is provided in the following subsections.

B. Listener selection

Listeners should be of ‘normal hearing’, and can be screened in advance through a calibration test. Voiers [3] mentions screening for self-consistency, which could also be accomplished, albeit more controversially, through rejection of extreme outliers in testing results. Otherwise, screening can assist in calibrating a signal-to-noise level of AWGN to yield a reasonable score for unprocessed speech. ANSI S3.2 [1] defines audiometrically normal as having hearing threshold levels no higher than 20dB and no lower than -10dB at any frequency from 125Hz to 8kHz as measured by an audiometer which complies with ANSI S3.6-1989 or ASA 81.

Where such audiometers are unavailable, a calibration test consisting of several CDRT words may well reveal gross hearing defects. An assumption may be made that frequency-selective hearing loss will be uncommon among volunteers for a hearing-based test, and in any case should be picked up by calibration testing. When using students for test subjects, it is useful to question subjects to identify any pattern of listening to loud rock music. In particular, subjects arriving at the test location wearing in-ear headphones at sufficient volume to be heard by the tester, are unlikely to make reliable test subjects. Voiers [3] considers eight listeners to be sufficient, however it is the practice of the author to typically arrange for 20 listeners.

During listener selection, it is important to consider and clarify confidentiality issues. It would seem that the optimum approach would be to maintain confidentiality by not identifying particular listeners by name. Results should not be made available to third parties in any way that could identify particular test subjects, without obtaining the prior agreement of the individuals concerned. It is also important to ensure that listeners are aware that this test does not reflect on their abilities in any way: it is not a hearing competition, but rather an assessment of the abilities of the system under test. Above all the test procedure should not in any way endanger the health of listeners, including any risk of hearing damage.

C. Test conduct

During the test, each listener is given a two-alternative forced choice of response for each randomly presented word. The choice consists of both words in the pair that contains the random word being replayed. For an automated push-button test, the word choice is normally presented prior to the sound playback, such that listeners are aware of the word choice before the word has been heard. Word choices may be presented in simplified or full character forms, hanyu pinyin plus tone, or any combination simultaneously.

For pencil-and-paper testing, the listener is always able to view alternatives prior to hearing the word. In either case, they are required to choose which word they have heard either through ticking a response sheet, or pressing the appropriate push-button. The latter method is advantageous in two ways: firstly in the ease of result collection, and secondly in the ability to self-pace the test (in that the next word should only be replayed once the current choice has been made). For pencil-and-paper testing, care must be taken to ensure that the listener expectation does not get out of sequence with respect to the audio playback.

Listeners should be familiarised with the test equipment and procedure, ideally through a trial or calibration run, prior to the main test. One useful technique is to prepend the last few words from the end of the test sequence to the beginning. These are then to be used for training and familiarisation, but are not counted as part of the results. It is important that the training words represent all of the speakers included in the test, and is often instructive to manually compare the results for repeated words. In theory, listener accuracy from the beginning will gradually improve through training, reach a plateau which should last for the duration of the test, before eventually reducing through listener fatigue. Fatigue occurs naturally, and is evidenced when tracking user scores over time through a drop in accuracy. This effect can be exacerbated by factors such as visual distraction, meaning that although tests do not need to be conducted in an anechoic chamber for auditory reasons, the sensory isolation available in such a location is a definite advantage.

In longer tests, fatigue can be combatted by allowing periodic rests during the test, with the trade-off that these rests will themselves increase test duration and may therefore be counter-productive. A two-speaker NCDRT test consisting of 160 basic words and 80 tone words, repeated twice, would result in 1920 words in the test corpus. If these are presented at 2 second intervals, test duration would exceed one hour, and listener fatigue effects could be expected. It is therefore preferable to conduct the NCDRT part I and part II tests on separate occasions. An optimal word presentation rate of once every 1.33 seconds has been reported by other authors [3] for the original English DRT test.

Both listener panels, and individual listening tests are possible, and for the NCDRT it is even possible to conduct the test individually using headphones. Where listener panels are
chosen, care should be taken to minimise distraction, and any possibility of copying.

D. Analysis of Results

The results post-processing advocated by Voiers [3] for the DRT should be used to eliminate the effects of guesswork. For $N_r$ correct responses and $N_u$ incorrect responses among $T$ test items, the percentage correct score, $P_c$, is obtained from (1):

$$P_c = 100 \frac{N_r - N_u}{T}$$

This score is applied to the overall NCDRT to provide a percentage intelligibility for all words within the test (for both reference and processed lists). The difference between these is a single intelligibility degradation metric for the entire system. NCDRT parts I and II scores can be calculated separately to identify phoneme and tonal intelligibility respectively. Within each part, a per-class result breakdown can be examined to determine a pattern to the degradation with respect to speech features of the classes. This can be useful in identifying aspects of the SUT which may require further attention or optimisation.

For NCDRT part I, the feature categories are shown in fig.3, and can in theory be related to the particular aspects of a speech handling system which cater for those features as explained below:

\{airflow-no airflow\} highlights the difference between consonants such as /p/ and /ch/ where airflow is large, to alternatives such as /b/ and /zh/ where airflow constriction takes place. Both lung excitation and pitch differ between the alternatives. \{nasal-oral\} underpins nasality testing where airflow and resonance through the nasal passage accompany vocal resonance, these are differentiated through pairs such as /m-b/ and /n-l/. However subjective experience indicates that many Chinese speakers will confuse consonant pair \{n-l\}. 

\{sustained-interrupted\} compares fricatives with stops or affricatives. “In Chinese, the fricatives are /f, h, x, sh, s, t/; the stops are /b, p, d, t, g, k/, and the affricates are /j, q, zh, ch, z, c/. Thus, the pairs of \{f-p\}, \{x-q\}, \{h-k\}, \{s-c\}, \{sh-ch\} and \{r-l\} are chosen (similar to the DRT except for \{r-l\} which is Chinese specific)” [4]. Since fricatives are generally of relatively high frequency, the fidelity of this class relates to the handling of the higher-frequency regions in speech, often reflecting on the quality of pitch handling in the SUT.

\{sibilated-unsibilated\} compares Chinese sibilants /z, c, s/ with /zh, ch, sh/ which are often and easily confused, again relating to fidelity of pitch handling.

\{grave-acute\} distinguishes words that are articulated through different tongue position at the front of the mouth, and are generally heavily voiced with low-frequency energy, thus this class can reflect on the handling of low frequencies by the SUT.

\{compact-diffuse\} are articulated behind the alveolar region of a mouth [4], again differentiated predominantly by tongue position.

NCDRT part II is better considered as an overall metric, rather than a per-category score. This is primarily because tone is related to pitch fidelity, meaning that a pitch coding unit would have responsibility for the score of all of the categories in the tone test. However, it may also be possible to categorise based upon the physical tone space differences in table III, in that some tones differ in absolute frequency, while others differ in change over time. Comparison of the degradation results for these two super-classes reveals the ability of the SUT to cater for absolute pitch location against its ability to convey a sufficiently fast pitch update rate.

E. Test Adaptation

The NCDRT adheres to the spirit of the DRT test, especially in the rhyming words of part I, whereas part II applies the same methodology to the measurement of tone confusion. Although the characters in the NCDRT have been chosen so as to be usable by the majority Chinese dialect group, in particular those with a Beijing-standard pronunciation, it is recognised that several regional variations in pronunciation of each character may exist. While the hanyu pinyin romanization does specify the phonetic characteristics, many Chinese, especially those of middle age or older, may not be comfortable reading hanyu pinyin and thus would need to be presented with characters during the execution of the test, therefore having few cues to pronunciation apart from recollection of the sound of each character. Since it is important to maintain the rhyming nature of word pairs, it is to be expected that users of the test may occasionally substitute different rhyming words into the test to match local pronunciations. In such cases the replacement words will preferably be similar to the originals in degree of voicing, sibilation, airflow, nasal resonance and CVC structure.

It is thus worthwhile repeating, and extending, the word selection principles from [4]:

1) Character pairs are to be used as the test unit, grouped by distinctive feature similar to the DRT. In NCDRT part I they occupy six phonetic groups, whereas in NCDRT part II they span all possible pairs of tone differences.

2) The monosyllable characters should take a CV or CVN form.

3) For NCDRT part I, only the initial consonant is different and tone should be the same across the pair.

4) For NCDRT part II, only the tone should differ - apart from this the pronunciation of both words in the pair should be identical.

5) Neither character in a pair should be hard to pronounce, and both should be in (ideally equally) common use.

6) The pronunciation of the characters should be single.

7) The pair list should cover as wide a syllable range as possible (different phonemes and tones).

VII. Conclusion

This paper has presented a New Chinese Diagnostic Rhyme Test (NCDRT) comprising two parts. Firstly a subjective phoneme intelligibility test based upon the English language diagnostic rhyme test adapted to a Chinese context. This test was largely evolved from the previously published Chinese Diagnostic Rhyme Test, but with some word replacement, and conversion to simplified jianti-zi form characters.
The NCDRT second part is a subjective tone intelligibility test, based upon new experimental evidence relating to the confusion of Chinese tone, also presented and analysed here. Together both parts of the NCDRT completely replace the previously published CDRT and CDRT-tone tests.

It should be noted that there is a large variability in published results of Chinese tone confusion. Issues regarding the use of carrier sentence (and thus the effects of tone sandhi), word choice, speaker or synthesiser used, methods of word degradation, amplitude or power normalization and subjective test methodology all affect results. In addition regional accents and dialects of test subjects play a part in overall DRT results. In this paper the use of a Beijing-standard announcer, large group of test subjects from across China, large word choice, and SNR-normalized word degradation were used to determine the confusion characteristics of DRT-style words spoken in isolation. This evidence has then been used to underpin the NCDRT methodology, in particular for the NCDRT part II tone tests.

This paper provide both parts of the NCDRT in hanyu pinyin romanized format plus tone identifier, and in simplified jianti-zi characters. Either can be used to conduct intelligibility testing, depending on listeners requirements. A complete procedural overview of the methodology for conducting the subjective Chinese intelligibility tests using the NCDRT is provided.

ACKNOWLEDGMENT

In reverse chronological order, the author wishes to acknowledge the contributions of Han Tingyue for her diligent testing of many volunteers, calibration trials and assistance with jianti-zi Chinese characters. Fong Loong Chong for his careful and thorough testing, particularly in pursuing different result patterns by Chinese heritage and dialect, and in-depth investigation of the curvilinear relationship between intelligibility and quality. Finally Dr Ding Zhong Qiang for establishing and validating the original tests and related follow-on work in pitch analysis in speech codecs.

Ian McLoughlin

Ian McLoughlin completed his PhD in audio signal processing from the School of Electronic & Electrical Engineering at the University of Birmingham in 1997 funded by Philips/Simoco Telecom in Cambridge. Prior to this he had worked for over 5 years first at the GEC Hirst Research Centre and later for Her Majesty’s Government. In 1998 he emigrated from the UK, spending almost four years in Singapore, then 5 years as Principal Engineer in Tait Electronics Group Research in Christchurch, New Zealand. In 2006 he returned to Singapore as Associate Professor at Nanyang Technological University School of Computer Engineering. He holds patents in speech intelligibility improvement and distributed ad-hoc wireless networking, and was co-recipient of the global 2005 IET Innovation in Engineering Award. He is also director of a fledgling electronics company, and a charitable trust supporting mission activities (http://www.missiontech.co.nz). He began learning Mandarin Chinese in 1991.

REFERENCES