Vowel intelligibility in Chinese

Ian McLoughlin

Abstract

Conventional wisdom states that, since the average amplitude of vowel articulation significantly exceeds that for consonants, an assessments of spoken intelligibility in obscuring noise should primarily be limited by consonant confusion. Furthermore, in both English and Chinese, consonant discrimination is considered to be more important to overall intelligibility than that of vowels. In the unbounded case, the assumption that vowel confusion is less important than consonant confusion may well be true, however at least two situations exist where the influence of vowel confusion may be greater. The first is where vocabulary-specific restrictions confine the structure of a particular spoken word to alternatives differing primarily in their vowel. The second is the prevalence of non AWGN interference, particularly impulsive noise which obscures only the vowel portion of a word, and similarly is present as a nonlinear effects of many time-sliced processing algorithms.

This paper explores the issue of vowel intelligibility for spoken Chinese, where the confusion characteristics are complicated through the influence of lexical tone carried by the vowel in CVC structure utterances. Experimental evidence from multi-listener intelligibility testing are presented to build toward an understanding of the characteristics of Mandarin Chinese vowel confusion in the presence of AWGN. Results are also isolated by carrier word consonants and in terms of the lexical tone overlaid upon tested vowels. In particular, several factors relating to issues such as vowel length, tone combination and the crucial influence of the /a/ (IPA [ɑ]) phone are revealed.

Index Terms

Mandarin, Chinese, intelligibility, tone, vowel, consonant

I. Introduction

SPEECH-based technologies have become increasingly important over recent years, not least through the near ubiquitous availability of wireless voice technology such as mobile phones. Significant social and economic wellbeing now depend upon quality speech communication over these networks, and as such any factors which reduce their quality are best minimised or eliminated altogether.

Coincident with the expanding role of wireless voice technology has been the economic rise of China as an emerging world superpower. Rates of cellular telephony ownership in China are high, and growing. In all likelihood, there will soon come a time - if not already passed - when the majority of worldwide speech processing is operating on Mandarin speech[1]. In commercial terms, Mandarin speech communication is likely to constitute the worlds largest and most important telecommunications market. For these reasons, special emphasis has been paid to the intelligibility of Chinese speech, in particular in relation to the aspects of speech affected by those communications systems.

Speech intelligibility assessment methods can be either subjective or objective. The former requires a group of human listeners, while the latter is typically conducted by automated systems. Tests can be made to evaluate either quality (how nice the speech sounds) or intelligibility (the ability to understand it). It is intelligibility evaluation which is the focus of the present paper: whilst perceived quality tends to sell systems, it is intelligibility which relates more closely to the ability to successfully conduct vocal communications.

Subjective intelligibility testing in Chinese has been performed using the proposed Chinese Diagnostic Rhyme Test (CDRT) standard for the past decade, enhanced with additions to assess tone discrimination, and the resulting evolution into a combined New CDRT (NCDRT) test methodology [1]. These tests, each based around part of ANSI standard S3.2 [2], have been applied to evaluate several speech coders such as GSM 06.10 [3] and ITU G.728 [4] for the conveyance of Mandarin speech.

The diagnostic rhyme test (DRT) A/B forced comparison method is one of the more popular intelligibility evaluation procedures enshrined in ANSI S3.2. This presents word pairs differing by a single attribute to listeners (see section III) [2], who are informed of two possible choices, and asked to select the correct one. Attributes which are mis-identified more often are classed as being more confused (or confusing) than others. Typically, two sets of word pairs are presented: one set has passed through a device under test (DUT), one has not. The difference between the confusion rates for attributes in each set is used to pinpoint the effect of the DUT on particular attributes. In the DRT, the differing attributes are simply the initial consonants from 96 rhyming word pairs. Published first by Voiers [5] in 1983, and used by the author for many years, the DRT demonstrates good repeatability and accuracy. In particular, this measure of intelligibility is considered a good predictor of overall speech intelligibility for a given system.

The NCDRT parallels the DRT methodology, but using Chinese words with a modified language-specific selection criteria. However the tonal nature of Chinese (see section II) which causes the understanding of Chinese words to be strongly dependent upon correct recognition of lexical tone, implies that measurement of consonant intelligibility alone is insufficient to predict

[1]The term ‘Mandarin’ is used to refer loosely to the majority Chinese dialect, and is used interchangeably with the term ‘Chinese’ in this paper.

overall intelligibility in Chinese. Thus the NCDRT specifically includes an additional corpus of test words for the express purpose of identifying tonal confusion as a second important predictor of overall intelligibility.

In general, in a DRT-style two alternative forced choice test, by selecting the attribute by which particular word pairs differ, the effect of these different attributes in speech can be assessed and evaluated on a relative scale through their differential confusion scores.

All DRT and NCDRT part I rhyming tests present words differing in a single attribute: their initial consonant. By varying the chosen consonants among word pairs, several speech attributes, such as sibilation or nasality, can be highlighted and investigated. In these and many other intelligibility tests, consonants are chosen for testing due to their relative importance to intelligibility, coupled with lower average spoken amplitude. Tempest, giving a detailed breakdown of measured spoken phoneme amplitudes (in [6] chapter 4), states that of 30 different spoken English phonemes, the nine vowels are spoken with highest average amplitude. In his tests, the /o/ in 'ford' was spoken with 28.3dB higher amplitude than the /th/ of 'thought'. In fact even the weakest vowel was 23.4dB above the /th/ reference.

In traditional voice communications circumstances, such as in POTS telephony, with obscuring noise being predominantly AWGN in nature, it is reasonable to assume that naturally quieter sounds will be confused more often and more easily than loud ones. Thus when computing overall intelligibility, emphasis should be paid to the quieter phonemes: consonants, particularly voiceless fricatives and plosives. Furthermore, in unconstrained English speech, consonants can be shown to convey a greater degree of intelligibility compared to vowels (see chapter 3 of [7]). This provides a further rationale for the emphasis on consonants in intelligibility testing.

Clearly no subjective test can exhaustively assess all phoneme confusions and combinations in a language. A listening test which lasts too long will invite listener fatigue and potentially be corrupted by learning effects. Thus pragmatic subjective testing methods must emphasise certain language characteristics over others. Given the constraints available, the choice of consonants for DRT and NDCRT testing has been reasonable.

However the author believes that vowel confusion also provides interesting and useful evidence which should not be ignored in intelligibility testing. There are several contributory reasons for this belief:

1) Modern speech systems no longer affect intelligibility predominantly through the action of AWGN. Digital erasure channels can exhibit periods of excellent SNR followed by total packet loss: lost or corrupted packets can just as easily affect vowels as they can consonants. Any test confined to consonant intelligibility would ignore the effects of something in the region of half of all random errors.
2) Some types of obscuring acoustic noise may also affect vowels and leave consonants relatively unchanged. For example noise consisting of repetitive impulses (such as that produced by many machines, or even by rainfall – water drops striking a resonant surface) is likely to mask speech more by virtue of when the impulses occur rather than the spectral or amplitude distributions of the noise. It is similarly not difficult to imagine that noise may exist having a spectral distribution more easily able to mask vowels rather than consonants.
3) In a CELP coder employing a perceptually weighted candidate scoring system, it is simplistic to assume that a loud phone, simply by virtue of being louder, will be less corrupted than a quieter phone.
4) Although vowels tend to be loud on average, glides, nasals and even the voiceless fricative /sh/ in 'ship' are similarly high in amplitude. Local variations in spoken amplitude may equally reduce the volume of particular vowels with respect to others (a 12dB variation within a sentence is not unusual).
5) Whilst consonants are highly important to English, they may be less so to other languages. The situation for Chinese in particular will be summarised in section II, where it is the vowel, not the consonant, which conveys a lexical tone. Otherwise, in a constrained-vocabulary situation, such as the NATO phonetic alphabet, word choices which differ primarily in their vowel will naturally depend upon vowel recognition for their intelligibility.

The hypothesis of this paper is thus firstly that vowel intelligibility is relatively independent of consonant intelligibility (in constrained CV[C] Chinese words), and secondly that significant variation exists in the intelligibility of vowels in CV[C] words despite the narrower range of spoken power across vowels compared to consonants. This provides evidence that the CV-structure NCDRT-style tests can, and in some cases should, investigate vowel intelligibility.

Experimental evidence will be gathered and analysed to support these hypotheses, in particular using the NCDRT testing methodology to invite comparison with consonant and tonal testing results. Chinese is particularly suitable for the language of experimentation due to its properties of economic and social importance, its CVC structure, and use of lexical tone on the vowel: however it is expected that many of the results will be similar for English and other languages.

Section II will begin by overviewing the relevant characteristics of Mandarin Chinese, section III will outline the test construction and implementation, section IV will analyse the test results and discuss the implications of this evidence, before section V concludes the paper.

II. Mandarin Chinese Speech

CHINESE is written as a sequence of unique stroke-based pictograms called characters. Every character is associated with at least one meaning and at least one pronunciation. In fact the alternative meanings for one character may be totally

TABLE I

HANYU PINYIN TRANSCRIPTION OF VOWELS AND CONSONANTS IN MANDARIN, WITH VOICED CONSONANTS SHOWN IN BOLD AND IPA REPRESENTATION GIVEN IN SQUARE BRACKETS. THE CONSONANT TABLE IS PRESENTED WITH ARTICULATORY INFORMATION FROM [9]

Consonants					
	Unspirated	*Aspirated*	*Nasal*	*Voiceless Fricative*	*Voiced Fricative*
Labial	b [b]	p [pʰ]	**m** [m]	f [f]	
Alveolar	d [t]	t [tʰ]	**n** [n]		**l** [l]
Velar	g [k]	k [kʰ]		h [x]	
Palatal	j [tɕ]	q [tɕʰ]		x [ɕ]	
Retroflex	zh [tʂ]	ch [tʂʰ]		sh [ʂ]	**r** [ʐ]
Dental Sibilant	z [ts]	c [tsʰ]	s [s]		
Finals consonants only:	**n** [n] **ng** [ŋ]				
Vowels					
a [ɑ]	e [ɣ]	i [i]	o [o]	u [u]	ü [y]
		ia [iɑ]		ua [uɑ]	
		ie [iɛ]			üe [yɛ]
ai [ɑi]	ei [ei]				
ao [ɑu]		io [io]		uo [uo]	
			ou [ou]		
uai [uɑi]	uei [uei]	iao [iɑu]	iou [iou]		

unrelated, particularly in the case of the simplified jianti-zi form characters used in the People's Republic of China (PRC), Singapore and elsewhere. Jianti-zi characters were originally derived from the traditional fanti-zi or full-form, characters used in Hong Kong and Taiwan, through a method of simplification that reduced the number of pen or brush strokes required for writing them. The simplification often resulted in totally different complex characters being reduced to identical simplified characters, however the meaning and pronunciation of the original was retained. Although two or more alternative pronunciations exist for many characters, the majority support only a single pronunciation.

As mentioned, each character has a particular meaning and pronunciation, and may exist alone, or be aggregated into compound strings which can alter its 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 [8].

A. The structure of Chinese speech

When spoken, all Chinese characters are monosyllabic. Multi-syllabic words are constructed from a string of characters. From this we consider the syllable as the basic unit of spoken Chinese [10]. Each of the character sounds can be written in the roman alphabet through the use of a hanyu pinyin romanization system [11] which will be used in this paper since it is in common use within China and is conveniently phonetic in nature. The Chinese Phonetic Alphabet (CPA) is similar, but includes several extra entries to account for the fact that the phonetics of hanyu pinyin are not totally invariant to context. Further information on the Chinese language and phonetics can be found in the analysis of DeFrancis [12]. Actually, the monosyllabic nature of Chinese confers a great advantage when performing subjective intelligibility tests compared to English: unlike English, all of the sounds presented to a Chinese listener can correspond to real words if so desired.

Furthermore spoken syllables adhere to a consonant-vowel-consonant (CVC) structure. Sometimes the initial consonant (C) is missing, and in many cases the final consonant is missing too, but the vowel (V) always remains. The final consonant, when present, is always nasal (N), being either /n/ or /ng/. Syllables thus conform to one of the combinations CVN, CV, V and VN, with about 415 permutations in use [10]. The consonants and vowels of Mandarin Chinese are listed in Table I (in fact many dialects of Chinese, including Hakka/Kejia, Cantonese/Guangdong-hua and Foochow/Fuzhou employ /ng/ in an initial as well as a final position, but Mandarin Chinese does not apart from the non-lexical use of /ng/ in isolation, typically as an interrogative). Note that some other phonemes exist, including the /er/ diminutive, very common as a final in Beijing Chinese (for the acoustics of Chinese vowels in the Beijing dialect see Zee and Lee [13]).

From a simple listing in Table I it can be seen that there exists a narrower range of voiced consonants in Chinese compared to English, a fact that would logically tend to increase the relative importance of these features to overall language intelligibility.

Given an estimated 13,000 written Chinese characters in use, and only 415 phonetic permutations, it follows that a very large number of homophonic characters exist. However the use of a lexical tone, applied to each and every spoken syllable, assists significantly in character differentiation [14].

B. Chinese tones

Mandarin has four distinct lexical tones plus a neutral, described as the lack of tone [15]. Applying a tone to a vowel means following a particular pitch contour during its utterance, with important criteria being the starting pitch, turning point (if any),

TABLE II

LEXICAL TONES IN CHINESE, THEIR IPA REPRESENTATION, CONTOUR, AND NUMERIC DESCRIPTION AND HANYU PINYIN REPRESENTATION

tone [IPA]	contour	numeric scheme	pinyin
1 [o͡]	high and steady	55 (53 before neutral)	ō
2 [o͡]	mid and rising tone	35	ó
3 [o͡]	mid dipping tone	214 (21 before neutral)	ǒ
4 [o͡]	high and falling tone	51	ò
neutral	context sensitive	-	o

gradient and ending pitch. Tone 1 has a relatively flat and consistent 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.

In order to more tightly define the tone, a numeric method was introduced in the 1930s (see [16]), where the pitch range is divided into five levels, with 1 being the lowest, 3 being mid-range and 5 the highest. This method has been less commonly referred to in recent times, but does neatly describe each tone by a sequence of pitch levels as shown in Table II. It is also important to note that the very common hanyu pinyin representation of tone is easily confused with the similar IPA symbols. Descriptions of tone such as 'mid' and 'high' are perceived rather than physically defined within given frequency ranges, it could be argued that these terms are to be understood in relation to mean pitch levels over an utterance.

The exact contour of the neutral tone is less important, having been described 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 [16]. Some tones in Table II also change slightly if preceding a neutral, which will in turn follow the adjusted endpoint of the preceding tone.

C. Chinese and intelligibility testing

The Chinese language is well-suited to DRT-style subjective intelligibility testing. Firstly the CVC nature of Chinese characters means that a regular test can easily be constructed using known characters. The sheer number of character choices for each phoneme combination allied with a careful selection procedure, ensures that the two characters in each pair are approximately as common as each other (see section III).

Although the role of consonants in Chinese intelligibility is arguably greater than that in English, the use of a vowel to carry the important lexical discriminator of tone, can be exploited during testing to either nullify or amplify the relative importance of the vowel by adjusting the role of tonal information used. The actual tonal and consonant confusions themselves, having already been established [1], the adjustment in the current paper is to minimise the effects of these features and thus obtain vowel confusion information.

III. VOWEL INTELLIGIBILITY EXPERIMENTS

A. Experimental design

A vowel confusion matrix, having 380 entries, can be obtained through experimentation whereby a word containing the vowel under test is spoken in the presence of noise, and listeners are required to indicate which vowel has been heard in an n-alternative forced choice test (and for Chinese $n = 20$). For example, 'bad' is spoken and listeners may respond with 'bed', 'bid' or similar. Over many repetitions of all vowels, a table of confusion can be created.

However there are certain difficulties with such a test. Firstly, to negate the effects of familiarity (when listeners may be more likely to 'hear' a common word than they are to hear a rare word [17]), each possible combination should have equal probability of recognition in the absence of other effects. Traditionally, many experiments have used nonsense syllables to obviate this issue. This is fine in English, however there are few unused combinations of CV words in Chinese: unless only a very limited set of CV combinations is explored, actual words must be used. For Chinese testing, the tone must also be considered – all of the possible alternatives should have common tone. Secondly, the requirement for a large confusion matrix would be for every vowel to be presented in the test, leading to an extensive test duration. Thirdly the effect of both tone and framing consonant(s) have to be taken into account. For example vowel confusion may differ when tested with certain initial consonants or tones.

There are two alternative approaches to accounting for the effect of tone and framing consonant(s). The first is to choose specific framing consonants and lexical tone and maintain these unchanged throughout the test [17] [18]. Both would need to be chosen carefully. The second approach is to use a spread of framing consonants and tones, and conduct sufficient tests to average out the effects of both.

This paper pursues the second approach: we will construct a test that aims to present each vowel multiple times, with a range of initial consonants and lexical tones [19]. This allows for investigation of the hypothesis that vowel intelligibility is, on average, independent of initial consonant intelligibility. Once this hypothesis has been proven, the effect of the initial consonant can be averaged out, the process repeated for lexical tone and the second hypothesis (that significant variation exists in vowel intelligibility) pursued.

TABLE III
DISTRIBUTION OF WORDS IN THE EXPERIMENTAL VOCABULARY

Criteria	Desired distribution	actual distribution
Ranged across 21 consonants	4 word pairs for each consonant	4 word pairs except for /q/ and /k/
Range of tone 1 words	25% (21 pairs)	26% (21 pairs)
Range of tone 2 words	25% (21 pairs)	18% (15 pairs)
Range of tone 3 words	25% (21 pairs)	26% (21 pairs)
Range of tone 4 words	25% (21 pairs)	30% (24 pairs)
Range of consonants	4 pairs	19×4, 1×3, 1×2
Range of vowels	4 pairs	7, 2×5, 3×4, 4×3, 13×2, 11×1

In Chinese, almost all initial consonants can be paired with every vowel and all, or nearly all, combinations exist with each of the lexical tones. This would provide good support for an n-alternative forced-choice test, such as the four-alternative forced choice approach in [19]. When selecting test words, it is quite apparent that not all combinations among the alternatives are equally common (e.g. there are 415 permutations of Chinese CVC words excluding the lexical tone - these can be found tabulated at the University of Vermont [20]). There is thus a need to select a reduced set of alternative words.

Since the ultimate aim of this testing is to provide evidence for the use of vowel-testing with an NCDRT structure, and a full range of alternatives can not be supported in any case, the multiple alternatives were collapsed to two, and a DRT-style two alternative forced-choice test constructed using CV permutations across the full range of tones, consonants and vowels.

B. Construction of the experiments

Having decided upon an NCDRT structure for the experiments, the critical issue is thus the choice of word pairs - or rather the choice of attributes to differ between word pairs. In this case, the decision was made to roughly adhere to the length of the DRT in order to minimise listener fatigue, with words evenly spread, as far as possible, across four main lexical tones and 21 initial consonants, amounting to 84 word pairs, within which each vowel was to be evenly represented. Furthermore, as mentioned previously, it was required that the relative frequency of each word in a pair, in normal speech, be approximately equivalent to minimise any skew due to unequal familiarity. In fact this approach was generally possible except in some cases where suitable word pairs could not be found, for example due to the lack of appropriate characters having initials /q/ and /k/. Table III identifies both the preferred even spread, and the actual spread of the resulting 81 word pairs.

The full list of characters used for the vowel intelligibility testing is shown in Fig. 1, giving the initial consonant, vowel and Chinese character for each pair, plus their tone. It can be seen from the list that the span of vowels covers several repetitions for each, matched with different consonants. The chosen words did not have any nasal finals and thus the terminating consonant of the CVC structure was null in each case - something which does not particularly limit the word choice but maintains consistency.

The aim when constructing the word list was to find three or four character-pairs for each initial consonant that spanned all of the vowels and tone combinations approximately equally. Ideally all characters were to be in reasonably common daily use, and each character in a pair should match the other in commonality. Where more than three word pairs could be found, preference was to be given to words having the least common alternative pronunciations and lexical tones. A panel of Chinese mother-tongue speakers undertook the character selection process and this was refined during the pre-test calibration procedures.

Characters were spoken by two announcers speaking standard Chinese: one male and one female. Both exhibiting clear and well-formed diction, but without a strong dialect (for example the /er/ final that is a feature of Beijing dialect was not present). The aim being to ensure that listeners, native Chinese speakers, were familiar and comfortable with the pronunciation and the words used [21]. Words were normalized by amplitude and had prefix and suffix consisting of at least 0.3 seconds of silence. Word presentation lists were randomised prior to any testing.

C. Calibration testing

Calibration tests used a total of six listeners, and were conducted in a similar manner to the final experimental tests. However each calibration testee repeated the entire procedure twice: once using white noise and once using rain noise. Equal numbers began with each type of noise, and then swapped over. It is interesting to note that the score difference between the two types of noise was less than 0.75%, and no evidence of learning effect between tests was observed: three listeners increased their score on the second test, and three decreased (net decrease was less than 0.5%).

All word pairs which exhibited either extremely poor or extremely good rates of recognition in the calibration tests were manually assessed to ensure accuracy. To further ensure quality, listener feedback was obtained regarding the test procedure, noise types, word choice, method of presentation and difficulty level. Apart from the decision to standardise on AWGN noise in the final experiment, no changes were made following the calibration tests.

	C	V	Char.	C	V	Char.	Tone		C	V	Char.	C	V	Char.	Tone		C	V	Char.	C	V	Char.	Tone
1	b	a	八	b	o	玻	1	28	h	ou	猴	h	u	湖	2	55	r	ao	绕	r	ou	柔	2
2	b	ai	百	b	ei	北	3	29	j	ie	节	j	üe	觉	2	56	r	u	入	r	uo	弱	4
3	b	i	必	b	ai	拜	4	30	j	i	几	j	u	举	3	57	r	uo	弱	r	ou	肉	4
4	b	ao	保	b	iao	表	3	31	j	ia	家	j	iao	交	1	58	s	uo	锁	s	i	死	3
5	c	ao	草	c	ai	彩	3	32	j	ie	借	j	i	记	4	59	s	ao	臊	s	ai	塞	1
6	c	e	厕	c	i	次	4	33	k	u	苦	k	ou	口	3	60	s	u	酥	s	uo	缩	1
7	c	a	擦	c	ai	猜	1	34	k	uai	快	k	ua	跨	4	61	s	i	四	s	e	色	4
8	c	u	醋	c	uo	错	4	35	l	ao	老	l	uo	裸	3	62	sh	a	沙	sh	i	湿	1
9	ch	a	叉	ch	ai	拆	1	36	l	iao	料	l	u	路	4	63	sh	u	树	sh	ou	受	4
10	ch	ai	柴	ch	ao	朝	2	37	l	ie	猎	l	e	乐	4	64	sh	ua	耍	sh	uai	甩	3
11	ch	e	车	ch	i	吃	1	38	l	ai	来	l	a	拉	2	65	sh	ao	烧	sh	uo	说	1
12	ch	ou	筹	ch	u	除	2	39	m	ao	冒	m	iao	庙	4	66	t	ou	偷	t	uo	脱	1
13	d	ao	岛	d	ou	抖	3	40	m	o	陌	m	u	木	4	67	t	ao	套	t	iao	跳	4
14	d	ai	呆	d	a	搭	1	41	m	i	迷	m	ei	玫	2	68	t	ie	贴	t	e	特	4
15	d	ou	斗	d	uo	朵	3	42	m	ai	买	m	ei	每	3	69	t	ai	台	t	u	图	2
16	d	i	敌	d	ie	碟	2	43	n	ai	奶	n	a	哪	3	70	x	iao	小	x	ie	写	3
17	f	ei	翡	f	ou	否	3	44	n	ei	内	n	i	逆	4	71	x	ia	下	x	iao	笑	4
18	f	ou	否	f	u	斧	3	45	n	ao	脑	n	iao	鸟	3	72	x	i	喜	x	u	许	3
19	f	o	佛	f	a	罚	2	46	n	a	拿	n	u	奴	2	73	x	u	需	x	ia	瞎	1
20	f	u	夫	f	a	发	1	47	p	o	破	p	ao	炮	4	74	z	ao	糟	z	uo	作	1
21	g	ai	该	g	e	哥	1	48	p	a	趴	p	ai	拍	1	75	z	ai	在	z	i	自	4
22	g	ou	狗	g	uo	果	3	49	p	o	婆	p	u	葡	2	76	z	e	责	z	u	卒	2
23	g	uai	怪	g	ua	挂	4	50	p	ei	配	p	a	怕	4	77	z	ou	走	z	uo	左	3
24	g	ao	高	g	u	骨	1	51	q	i	起	q	u	取	3	78	zh	a	眨	zh	i	纸	3
25	h	ua	画	h	uai	坏	4	52	q	ie	切	q	iao	敲	1	79	zh	e	这	zh	u	住	4
26	h	a	哈	h	ua	花	1	53	q	ie	切	q	üe	缺	1	80	zh	ai	宅	zh	a	炸	2
27	h	ei	黑	h	ai	嗨	1	54	r	i	日	r	e	热	4	81	zh	uo	桌	zh	ua	抓	1

Fig. 1. List of characters used in the vowel intelligibility experiments

D. Experimental procedure

The final experiments were performed in a noise-free and distraction-free environment. Tests were performed using a computerised system that presented words (in both Chinese character form and hanyu pinyin as per [19] and [1]) to the listener approximately 0.5 seconds prior to audio playback over headphones. Due to the silent prefix to each recorded character, additive noise began 0.3 seconds prior to playback of each word. Subsequent replay of presented sounds was disallowed, and the self-paced test deliberately did not have options to go back to, or correct, previous answers. The two alternatives were forced: there was no provision for a 'don't know' answer. The system collected test statistics automatically.

The main experiment comprised 20 male and female listeners aged between 18 and 30. Listeners were pre-selected using two major criteria: the first being normal hearing, and the second being a level of familiarity with written and spoken Chinese to ensure that each of the characters used in the tests would be recognisable by them. Listeners, unpaid volunteers, were all mother tongue speakers of Chinese. They were introduced to the system through a familarisation session before the procedure began in an attempt to reduce training effects during the test.

The noise to speech SNR was adjusted based upon listener confusion rates (during calibration). The intention was to avoid a response score near to limits of either 100% or 50% since these extremes correspond respectively to perfect intelligibility, and to such poor intelligibility that the result score is indistinguishable from guesswork.

IV. Experimental Results

Overall the level of AWGN obscuring the speech was such that 63.6% of vowels were correctly recognised, after guesswork elimination [1] with Fig. 2 showing the rates of confusion between tested vowels. Due to the large size of the confusion matrix between all possible vowel pairs (400), related to the test scope (162 words), many potential vowel pairs were untested following the test selection procedure, concentrating instead upon the likely most confused subset of the entire matrix. This approach mirrors that of other authors working in Chinese [19] and English [18] [17] [22].

The most confused pairs of vowels were [ie –uë], [ua – uai], [i – u], [ei – ou] and [a – ia]. The least confused formed a long tail beginning with [ia – u] and [ai – e].

An interesting relationship evident in these results is in the influence of the phone /a/ toward vowel confusion. When ordering vowel pairs in order of confusion, the bottom $1/3$ of the list (i.e. the least confused pairs) included the phone /a/ in one half of the pair. By contrast, the $1/3$ *most* confused vowel pairs either did not include the /a/ or included /a/ in both vowels, i.e. the /a/ sound was not used for differentiation in those cases. It thus seems that in general, vowels containing /a/ are unlikely to be confused for a vowels that do contain /a/ and vice versa. For example, word pairs such as [jia, jiu] are unlikely to be

confused because /a/ appears in only one of the alternatives, whereas [jia, jiao] are more likely to be confused because /a/ appears in both alternatives.

The experimental results can be presented according to the number of times words beginning with particular consonants were confused. This is still evaluating vowel confusion, but is considering whether the consonant before a particular vowel plays any part in the likelihood that the vowel following it is confused. In effect, this is investigating the nature of the framing syllable. Fig. 3 plots a histogram of the percentage of times that vowels framed by the named initial consonant were confused. Clearly, vowels with initials /j/, /n/ and /q/ were confused more often than others: two are palatals, one is alveolar. Although these were, on average, highly confused, out of the nine word pairs having these initials, only four were from the list of most confused vowels. By contrast, /b/, /z/, /f/ and /p/ initials were least confused: three are labials, one is a dental sibilant (although /z/ is often voiced). Note that, as expected, both the absolute percentage error attributed to initial consonant ($< 9\%$) and the spread ($< 7\%$) are significantly less than the figures obtained from the vowel contribution.

As a check on any undue influence by particularly erroneous word pairs, Fig. 4 plots the number of listeners who were confused by the same word pairs. The majority of the incorrectly recognised word pairs were only confused by a few of the listeners. In fact none of the word pairs was confused by all listeners, and only two of the 81 word pairs was confused by a majority of listeners.

At this point, it is possible to proceed with the testing of the first hypothesis in Section I that vowel intelligibility is relative independent of the initial consonant. This would be in line with a probability-based quantitative approach [19], where the recognition probability of a phoneme, P_p is related to the multiplicative recognition probabilities of the vowels P_v, consonants P_c and tones P_t in eqn. 1:

$$P_p = P_c^{w_c} P_v^{w_v} P_t^{w_t} \tag{1}$$

where w_c, w_v and w_t are weighting parameters to account for the different importance of those features to overall intelligibility (recognition probability). There is some evidence to suggest roughly equal weighting factors in Chinese for CV phonemes [19].

In the present experiment, the assessment of the correlation between the probability of vowel and that of consonant was performed using a Spearman ranking. In this case, words were ranked in terms of errors attributed to the vowel and errors attributed to the consonant. The Spearman correlation coefficient ρ was obtained from the difference, d_i between the $n = 160$ non-zero rankings of the experimental results as in eqn. (2):

$$\rho = 1 - \frac{6 \times \sum_{i=0}^{N} d_i^2}{n^3 - n^2} \tag{2}$$

Yielding a ρ of only 0.25, there is thus only a very weak correspondence between error rates, vowel confusion depends primarily upon the vowel-pair under consideration, rather than upon the initial consonant. This supports the first hypothesis mentioned in Section I. The second hypothesis, that significant variation can be seen in vowel intelligibility, is also supported by the results, in particular those of Fig. 2 where the error percentage ranges from zero to over 50% for different vowel pairs. It should be noted that the across-phoneme averaging implied in this analysis may hide the existence of certain specific vowel-to-consonant dependencies. These could be revealed using an experimental design which fixed a subset of consonants and tones for vowel testing, however this would then lose the wider focus necessary for unconstrained phoneme combinations, of the first hypothesis.

The influence of tone on vowel intelligibility can be summarised by considering the average confusion rate for words carrying the four tones: vowels carrying tone 4 were the most confused overall; in 23% of cases, compared to tone 3 (20%), tone 1 (14%) and tone 2 (13%). Although the reason for this has not yet been isolated, it is probable that the steeper gradients of tones 4 and 3 play a part in obscuring the identity of the underlying vowel. Whilst there may be no meaningful significance to the results between tones 3 and 4, and likewise between the results for tones 1 and 2, it is significant that tones 3 and 4 lend themselves to greater confusion levels than tones 1 and 2. Interestingly, dedicated tone intelligibility tests find that, in the absence of dependencies due to tone sandhi, tones 3 and 4 are least confused for other tones [19] [1]. The current results are dissimilar in that they are not highlighting situations where one tone is confused for another, but rather where tonal shape, when applied to a vowel, lends itself to greater levels of vowel confusion (not tonal confusion). It thus appears that the *least* confusing tone(s) plays a part in reducing the intelligibility of the vowel which carries it. The probable interpretation is that the very aspect of tone that contributes to its intelligibility, namely the greater tonal modulation – in both time and frequency [23] – acts to distort the intelligibility of the vowel which carries it.

One final result from the testing was that two listeners reported that they considered that the length of the spoken vowel was occasionally the prime differentiating factor between two alternative vowels: even without recognising the sound of the vowel itself, they felt it was possible to ascribe the choice correctly by listening to the length of the voiced section of the word. It is possible, albeit requiring further experimentation to clarify, that a similar process is at play during the operation of the McGurk effect [24], at least in those phonemes not obviously distinguished through lip and visible tongue position.

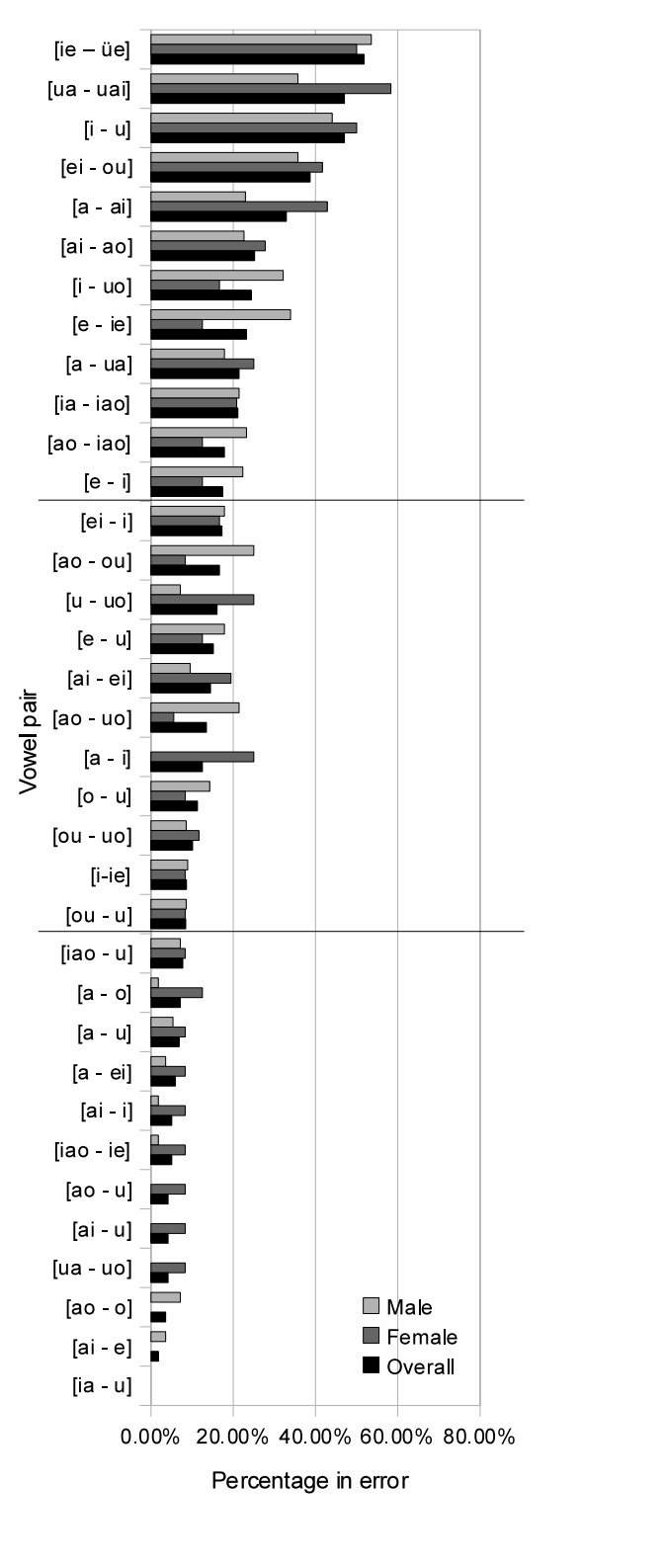

Fig. 2. A histogram showing the percentage of misidentified vowel pairs for the male speaker, the female speaker and overall after guesswork elimination has been applied to the results [1].

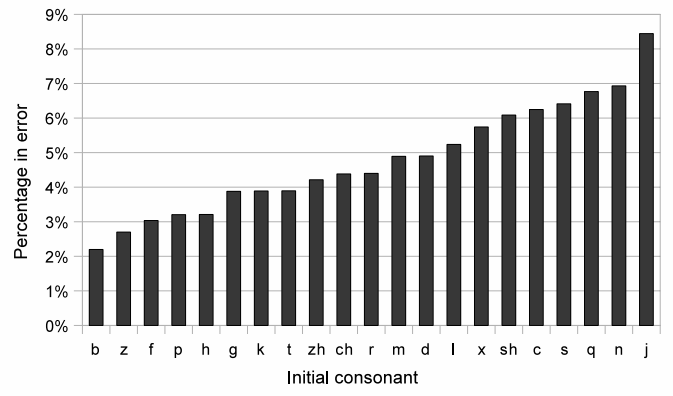

Fig. 3. A histogram of the percentage of times word pairs beginning with each of the given consonants were confused.

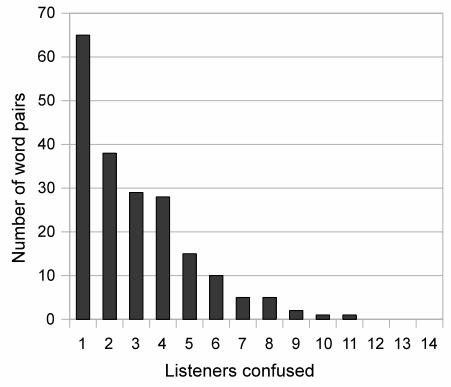

Fig. 4. A histogram showing the distribution of word pairs by the number of listeners who confused them.

To determine whether vowel length did in fact play a part in the test, the duration difference was measured between the vowels of each tested vowel pair. To ensure accuracy, this tedious measurement for all words was undertaken manually by an expert.

The results, presented in Fig. 5, plot the percentage difference in vowel duration between the two vowels in the word pair, versus the number of incorrect answers attributed to that pair. The six highest and lowest duration difference values ($< 1\%$ and $> 50\%$) were excluded from the plot for clarity. Over the remaining sample values, averaged over more than 2000 data points, there is little evidence that the difference in vowel duration contributes to confusion. However note the gradient of the trend line in Fig. 5: words with greater difference in duration actually have a tendency to be mistaken more often. This is probably not indicating a general relationship. Further analysis confirms that the confusion does not reduce with the increase in the absolute spoken duration difference of vowels. In fact there is a slight correlation in the opposite direction: a weak relationship appears to exist between confusion rate and duration difference (when errors per vowel pair, and different in spoken duration, are quantized into bins of 10%, an exponential fit, which is slightly better than either linear or quadratic, yields a fairly poor correlation coefficient R^2 of 0.66).

Finally, there was some evidence of vowel confusion, among the simple vowels at least, following a vowel loop [17], where

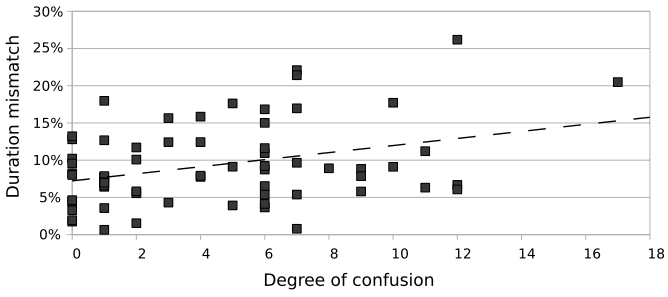

Fig. 5. A scatterplot of word pair duration difference against number of confusions, overlaid with a linear fit trend line.

individual vowels on the loop were most often confused for their neighbours. For example, /i/ [i] was most often confused for /u/ [u] (44%), then for /e/ [ɣ] (27%). However there were two significant deviations from a straightforward vowel loop, namely the confusion of /e/ [ɣ] for /u/ [u] (18%) and /a/ [ɑ] for /u/ [u] (11%).

V. Conclusion

This paper has discussed the role of the vowel in the intelligibility of Chinese CV structure words. It has assessed the confusion between different vowels through constructing and running a listening experiment using DRT-style two-alternative forced-choice word pairs. Hanyu pinyin vowels were presented randomly, framed with a succession of initial consonants, overlaid with a selection of tones, and presented to listeners in high levels of AWGN to identify the contribution to word intelligibility due to each vowel.

Results identified the most and least confused vowels, the contribution of initial consonant choice to vowel intelligibility, and the central role of the /a/ phone in vowel, and word, intelligibility for Mandarin Chinese. Overall, vowel intelligibility was found to be relatively insensitive to initial consonant choice, and vowel intelligibility itself exhibited significant variation across the range of Mandarin Chinese vowels.

The contribution of tone to vowel intelligibility was also explored, showing that tones 3 and 4 led to lower vowel intelligibility than tones 2 and 1, although when tested in isolation, tones 3 and 4 are typically more intelligible than tones 1 and 2. Finally the issue of vowel duration, reported by some listeners as being important to their own perception during listening tests, was considered and found to not be a generalised result for discrimination of Mandarin in noise.

Acknowledgment

The author wishes to acknowledge the contributions of Han Tingyue, Lim Dawai and Guo Zhengjin for assisting with Mandarin listening tests, and Farzane Ahmadi for experimental oversight to ensure quality and repeatability.

Ian McLoughlin completed his PhD in audio signal processing at the University of Birmingham, UK in 1997 funded by Philips/Simoco Telecom, Cambridge. Prior to this he had worked first at the GEC Hirst Research Centre and later for Her Majesty's Government. In 1998 he began almost four years lecturing 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 has been working in the field of Mandarin Chinese speech for around 17 years.

REFERENCES

[1] I. V. McLoughlin, "Subjective intelligibility testing of Chinese speech," *IEEE Trans. Audio Speech and Lang. Proc.*, vol. 16, p. 2333, Jan. 2008.

[2] ANSI, "ANSI S3.2," *Method for Measuring the Intelligibility of Speech over Communication Systems*, 1989.

[3] I. V. McLoughlin, Z. Q. Ding, and E. C. Tan, "Evaluation of the GSM speech coder using the proposed Chinese diagnostic rhyme test speech intelligibility measure," *Speech Communication*, vol. 38, pp. 161–165, 2002.

[4] F. Chong, I. V. McLoughlin, and K. Pawliowski, "Evaluation of the ITU-T G.728 as a voice over IP codec for Chinese speech," in *Australian Telecomms. and Networking Apps. Conf.*, Melbourne, Dec. 2003.

[5] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, pp. 30–39, Jan. 1983.

[6] W. Tempest, Ed., *The Noise Handbook*. Academic Press, 1985.

[7] I. V. McLoughlin, *Applied Speech and Audio Processing*. Cambridge University Press, 2009.

[8] F. Coulmas, *The Blackwell encyclopedia of writing systems*. Blackwell Publishing, 1999.

[9] P. H. Zein. (2009) Mandarin Chinese phonetics. [Online]. Available: http://www.zein.se/patrick/chinen8p.html

[10] Z. Li, E. C. Tan, I. McLoughlin, and T. T. Teo, "Proposal of standards for intelligibility tests of Chinese speech," *IEE Proc. Vision Image and Sig. Proc.*, vol. 147, no. 3, pp. 254–260, June 2000.

[11] "HanYu PinYin FangAn," *First National People's Representatives Meeting, 5th Conference*, 1958.

[12] J. DeFrancis, *The Chinese language: Fact and fantasy*. University of Hawaii Press, 1986.

[13] E. Zee and W.-S. Lee, "An acoustical analysis of the vowels in Beijing Mandarin," in *EUROSPEECH*, 2001, pp. 643–646.

[14] J. M. Howie, *Acoustical Studies of Mandarin Vowels and Tones*. Cambridge University Press, 1976.

[15] M. Yip, *Tone*. Cambridge University Press, 2002.

[16] C. N. Lee and S. A. Thompson, "The acquisition of tone in Mandarin-speaking children," *Journal of Child Language*, vol. 4, pp. 185–199, 1979.

[17] G. Peterson and H. Barney, "Control methods used in a study of the vowels," *J. Acoustical Soc. America*, vol. 24, no. 2, pp. 175–184, 1952.

[18] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoustical Soc. America*, vol. 97, no. 5, pp. 3099–3111, 1995.

[19] Q.-J. Fu, F.-G. Zeng, R. V. Shannon, and S. D. Soli, "Importance of tonal envelope cues in Chinese speech recognition," *J. Acoustical Soc. America*, vol. 104, no. 1, pp. 505–510, 1998.

[20] J. J. hua Yin. (1998, Aug.) Chinese phonetics. [Online]. Available: http://www.uvm.edu/ chinese/pinyin.htm

[21] T. L. Gottfried and T. L. Suiter, "Effect of linguistic experience on the identification of Mandarin Chinese vowels and tones," *Journal of Phonetics*, vol. 25, pp. 207–231, Apr. 1997.

[22] A. Cutler, A. Weber, R. Smits, and N. Cooper, "Patterns of English phoneme confusions by native and non-native listeners," *J. Acoustical Soc. America*, vol. 116, no. 6, pp. 3668–3678, 2004.

[23] X. S. Shen, M. Lin, and J. Yan, "F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3," *Journal of the Acoustical Society of America*, vol. 93, pp. 2241–2243, Apr. 1993.

[24] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, p. 746748, 1976.