The International
Journal of
Speech,
Language
and the Law

IJSLL (PRINT) ISSN 1748-8885
IJSLL (ONLINE) ISSN 1748-8893

Article

Bilingual speaker identification: Chinese and English

Peggy P. K. Mok, Robert Bo Xu and Donghui Zuo

Abstract

Very few studies have examined voice memory and speaker identification in bilingual contexts. This study investigates how well bilingual listeners can identify bilingual voices in different language conditions. 89 Cantonese-English and 89 Mandarin-English listeners participated in voice line-ups with Cantonese-English voices in the same-language and cross-language conditions. Results show that the overall identification accuracy is low. Cantonese-English listeners perform significantly better in the same-language than cross-language conditions, similar to previous findings based on monolingual subjects. However, there is no language effect for the Mandarin-English listeners, possibly due to their unfamiliarity with the languages concerned. Confidence ratings show that all listeners are more confident in the same-language condition with their most familiar language, although the relationship between confidence and accuracy is not reliable. The results suggest that some indexical information about speaker identity is language-dependent. Different articulatory settings may explain the better performance of Cantonese-English listeners in the same-language conditions.

KEYWORDS SPEAKER IDENTIFICATION, VOICE MEMORY, BILINGUAL, CANTONESE, ENGLISH

Affiliation
The Chinese University of Hong Kong
email: peggymok@cuhk.edu.hk robertbaldwinxu@gmail.com donghuizuo@gmail.com

IJSLL VOL 22.1 2015 57–77
doi : 10.1558/ijsll.v22i1.18636

equinox www.equinoxpub.com

1. Introduction

Remembering and recognising the voice of a particular person, something that most of us do almost on a daily basis, is an important cognitive ability and social skill. We all have experience in successfully recognising people by their voices alone, but incorrect identification also occurs, even for familiar voices (Foulkes and Barron 2000). In addition to being socially relevant, speaker identification is also a useful, sometimes critical, tool in court cases of various sorts, when an earwitness is available. Earwitness identification involves recognition of voices heard at a crime scene by untrained listeners. The accuracy of voice recognition and speaker identification is influenced by many factors, e.g., speaker familiarity, listening conditions, duration of voice heard, and telephone transmission (Hammersley and Read 1996; Yarmey 1995, 2007). However, only a few studies have investigated the influence of language background on speaker identification, and even fewer have assessed the effects of bilingualism. The present study aims to fill the gap in this under-studied area.

Two types of properties are conveyed simultaneously in the speech signal: linguistic and indexical (Abercrombie 1967). Linguistic properties refer to the message a speaker is trying to convey (what is said), while indexical properties refer to the extralinguistic cues signalling personal characteristics of a speaker, e.g., age and sex (how it is said). Speaker identification involves both types of properties, but the interaction between them remains poorly understood. Mixed results were reported regarding the (in)dependence of linguistic and indexical properties in speech perception (see reviews in Stockmal, Moates and Bond 2000 and Winters, Levi and Pisoni 2008). Investigating the identification of bilingual speakers across languages provides an opportunity to evaluate the importance and interaction of these two types of properties in speech processing.

There are also practical reasons for studying bilingual speaker identification. Language contact and bilingualism[1] is very common in many parts of the world (Grosjean 2013), but most previous studies on speaker identification were based on monolingual speakers and listeners. Moreover, more and more legal cases occur in which a lay earwitness has to deal with speech material from a language that s/he does not understand or speak natively (e.g., Köster and Schiller 1997; Rogers 1998). The reliability of their speaker-identification ability is in question. Therefore, it is important to investigate how bilingualism affects speaker identification by lay witnesses using both bilingual speakers and bilingual listeners for a better understanding of these issues.

1.1 Influence of language background on speaker identification by lay witnesses

Only a few studies have examined the relationship between language background and speaker recognition. Goldstein, Knight, Bailis and Conover (1981) tested monolingual American English listeners' recognition of unfamiliar voices under different conditions: voices with and without accents (African American English and Chinese) and voices speaking in a foreign language (Spanish). They found that recognition of voices speaking an unknown language was no worse than recognition of accented voices, and that there was no difference between voice recognition of foreign voices and native voices, which means that there was no effect of language background in their study.

However, other studies showed that language background does affect voice identification. Thompson (1987) had monolingual English listeners hear voices of bilingual students speaking English, Spanish, and English with a strong Spanish accent, and identify those voices in line-ups. There was a clear effect of language: voices were identified best when speaking English and worst when speaking Spanish, with recognition accuracy of the accented voices intermediate between the two. Goggin, Thompson, Strube and Simental (1991) also showed that language familiarity is an important factor in speaker identification. Monolingual English and monolingual German listeners were better in identifying bilingual voices speaking in their native languages, respectively. Monolingual English listeners in their study showed the same order of identification accuracy as in Thompson (1987). Köster, Schiller and Künzel (1995) tested three groups of listeners differing in their knowledge of German (native German listeners, English listeners who learned German, and native English listeners with no knowledge of German). They found that listeners with knowledge of German generally performed better than those without any knowledge of German in recognising a German speaker, but there was no difference between German native speakers and English learners. The native language advantage is also found in Philippon, Cherryman, Bull and Vrij (2007) using English and French material for English listeners. Moreover, the language effect can be extended to dialects and accents as well. Sjöström, Eriksson, Zetterholm and Sullivan (2008) found that voice identification in a familiar accent was more accurate than in a less-familiar accent.

In addition, studies by Köster and Schiller (1997) and Wester (2012) showed that typological difference of the target language from listeners' native language did not impact their identification. Again, knowledge of the target language helped listeners in voice identification, but as long as the target voice is in a foreign language, what language it is did not further affect the results of identification in their studies.

The aforementioned studies demonstrate clearly that language familiarity has an influence on speaker identification. Voices speaking the native languages of the listeners are easiest to recognise, followed by voices speaking the languages with an accent. Voices in an unknown language are the most difficult to recognise. A logical question arises: is linguistic information in the speech signal an important factor in the identification process? Goggin et al. (1991) showed that, when the passage being spoken is made increasingly less similar to English by rearranging words, rearranging syllables and reversing text, English-dominant listeners' voice recognition deteriorated systematically across these conditions. Schiller, Köster and Duckworth (1997) removed linguistic information by replacing all syllables in a natural German passage with 'ma' in order to minimize the linguistic cues to the target language, while keeping some phonetic and phonological features. They found that German and English monolingual listeners as well as English listeners with some knowledge of German did not differ significantly in their recognition ability. Thus, these two studies show that linguistic information is essential in speaker identification.

Besides identification accuracy, the confidence level of the listeners is also affected by language familiarity. Listeners seem to have a greater confidence towards a familiar language in the recognition tasks, even though the correlation between accuracy and confidence is not reliable (Goggin et al. 1991; Hammersley and Read 1996; Sørensen 2012; Thompson 1987; Yarmey 1995, 2004). Using speakers and listeners of the same language, Yarmey, Yarmey, Yarmey and Parliament (2001) found that the correlation between accuracy and confidence is stronger for highly familiar speakers than unfamiliar speakers.

1.2 Identification of bilingual speakers

Bilingualism adds an interesting perspective to the discussion of language background in speaker identification by lay witnesses. However, only very few studies have investigated this topic. To the best of our knowledge, so far, only one study involved both bilingual speakers and bilingual listeners. Goggin et al. (1991) found that English-Spanish bilingual listeners were less affected by the language condition in their experiment than the monolingual English listeners were for both confidence and accuracy, showing that bilingual listeners performed similarly in both languages. However, their listeners heard the same language condition in both the initial exposure and in the identification task, i.e., there was no 'crossing over' of languages in the experiment. Therefore, it is still unclear if bilingual listeners can identify bilingual speakers based on the memory of their voices in one language and identify them in another language.

Two recent studies on the identification of bilingual speakers by monolingual listeners in the form of same-language and cross-language pairs can provide some insights into the above question. Winters et al. (2008) showed that native monolingual English listeners could discriminate voices of German-English bilingual speakers. They performed better with voices speaking in a familiar language, and also performed better in the matched-language than in the mixed-language condition, but some listeners could still transfer some knowledge of the speakers' voices from one language to the other language. Their results show that there is sufficient language-independent information in speech to make identification of speakers across languages possible. Wester (2012) extended their findings to more language pairs: German-English, Finnish-English and Mandarin-English. She also found that monolingual English listeners could discriminate the voices of bilingual speakers, but they again performed much better in the matched-language than in the mixed-language condition. Additionally, there was no significant difference in all matched-language conditions, regardless of whether it was in English or a foreign language.

Results in the above two studies show that monolingual listeners can generalise some knowledge of speakers' voices across languages. The mismatch of linguistic and indexical information in the mixed-language condition may have increased the cognitive load, which can explain the difficulties faced by monolingual listeners. They appeared to process indexical information in a language-dependent way when they could understand the language. Indexical information is processed in a language-independent manner when linguistic information is lacking. However, it is still unclear if these findings based on monolingual listeners are equally applicable to bilingual listeners who have access to both indexical and linguistic information simultaneously.

Stockmal et al. (2000) examined the importance of linguistic and indexical information from a different perspective. They used bilingual speech material produced by the same talkers in language pairs all unknown to monolingual English listeners in same-language/different-language conditions, thus controlling the indexical information. They found that listeners could discriminate between foreign languages even when the material was produced by the same bilingual speaker, although not all language pairs were equally discriminable. There is sufficient phonetic information in the speech signal for listeners to tell unknown languages apart. Thus, their results show the separation of linguistic and indexical information.

1.3 The present study

This study investigates speaker identification by lay witnesses using both bilingual speakers and bilingual listeners. As mentioned above, only one previous study involved both bilingual speakers and listeners (Goggin et al. 1991) but there was no 'crossing over' of languages in their experiment. Although Winters et al. (2008) and Wester (2012) included the mixed-language condition in their experiments, their monolingual listeners could only understand the content of one language, and thus showed a strong native language advantage in their identification. As bilingual listeners can understand the content in both languages, it is unclear whether and how the native-language advantage will affect bilingual listeners. Moreover, both Winters et al. (2008) and Wester (2012) used discrimination tasks in their experiment, not voice line-ups that are usually employed in forensic speaker identification (e.g., Sullivan and Schlichting 2000; Yarmey 2001). Most previous studies on speaker identification also told the listeners to remember the voice explicitly during exposure, which does not simulate 'real-life' forensic situations where memory of voices is important in identifying a speaker. In order to have a more thorough understanding of bilingual speaker identification, the present study has a novel approach in studying speaker identification: having bilingual listeners identifying voices of bilingual speakers in voice line-ups in two language conditions: same-language and cross-language. Two groups of bilingual listeners were included: Cantonese-English listeners who understand the speech material in both languages (Experiment 1) and Mandarin-English listeners who only understand the material in English (Experiment 2). Comparing these two groups of listeners allows us to assess the native-language advantage in bilingual listeners. Moreover, we tested listeners' unexpected memory of voices by not instructing them to remember the voice during exposure. Finally, this study also investigates how confident bilingual listeners are when they make judgments in different language conditions, and how reliable their confidence is.

2. Experiment 1: Cantonese-English

2.1 Method

2.1.1 Voices and materials

Ten female native speakers of Cantonese who spoke English fluently with a Cantonese accent were invited to the recording session, which took place in a sound-treated recording booth at the Chinese University of Hong Kong. All of them were students at the Chinese University of Hong Kong, and their voices were judged to be common and 'unmarked' impressionistically (not having any distinct personal characteristics, e.g., hoarseness). They were asked in Cantonese to give a long spontaneous answer to each of the four questions (two in Cantonese and two in English). For the Cantonese questions, the speakers were asked to describe: a) a travel experience, and b) a dining experience. For the English questions, the subjects were asked to talk about: a) their study at the university, and b) their study in secondary school. The speech material was recorded directly onto hard disk using Praat with a sampling rate of 22050 Hz via a condenser microphone placed approximately 20 cm away from their mouths. The recordings were auditorily screened by the authors, and three voices with some identifiable features in the recordings were excluded. The remaining seven female voices were judged not to have any obvious idiosyncratic characteristics, and have a similar style (speech rate and accent). These voices were used in the identification experiments. Using auditory analysis, three of the voices were judged to be highly similar by the authors. One of the similar voices was chosen as the target voice and the other two were used as 'distractors' (Foil A and Foil B). Table 1 shows the average articulation rate and fundamental frequency (F0) of all speakers. Articulation rate was calculated based on the average number of phonological syllables per second in the 30-second samples used in the line-ups (more details below). Pauses and hesitations were eliminated when preparing the 30-second samples. Average F0 was based on the F0 tracked every 200 ms of the 30-second samples automatically by Praat. It can be seen that small differences exist even between the exposure voice and the target voice used in the line-ups, both of which came from the same chosen speaker. All speakers spoke faster in Cantonese than in English. Most of them also had a slightly higher F0 when speaking in English than in Cantonese. Foil A has a very similar articulation rate and F0 values to the target voice.

The English (university study) and Cantonese (dining experience) speech material of the target voice was shortened to two one-minute samples. These samples were used as familiarisation material for the listeners during the exposure part of the experiment. The speech material in English (secondary study) and Cantonese (travel experience) by all speakers was cut into 30-second samples. To minimise the effect of context on identification, the orders of the utterances in these 30-second samples were randomised within each narrative, and the randomised versions were used in the voice line-ups. Pauses, hesitations and any obvious idiosyncratic vocalisations (e.g., cough) were removed.

Table 1: Average articulation rate and fundamental frequency of the speakers

	Articulation rate (syllables/second)		Fundamental frequency in Hz (SD)	
Voices[a]	Cantonese	English	Cantonese	English
Exposure	4.40	3.85	173 (36.0)	192 (40.1)
Target	4.75	3.30	180 (53.6)	193 (27.8)
Foil A	4.58	3.31	195 (32.8)	195 (30.3)
Foil B	3.95	3.28	201 (41.8)	225 (48.4)
Foil C	5.51	3.66	190 (33.9)	199 (38.0)
Foil D	5.02	4.06	224 (47.6)	230 (35.1)
Foil E	3.86	3.05	237 (60.5)	262 (44.7)
Foil F	4.99	3.59	200 (33.4)	215 (34.3)

[a] Exposure and Target voices came from different recordings of the same speaker.

2.1.2 Listeners

Eighty-nine Cantonese-English bilingual listeners participated in the identification experiment. They were undergraduate or MA students at the Chinese University of Hong Kong, between 18 and 29 years old (mean = 21.46, SD = 2.70).

The bilingual listeners all grew up in Hong Kong and spoke Cantonese as their native language. They started learning English before the age of six. All of them could communicate effectively in English, although they used English more in academic settings. None of them reported any history of hearing or speech problems. All listeners received course credits for their participation.

2.1.3 Procedure

The 89 bilingual listeners were randomly assigned to one of the four language conditions: Cantonese exposure and Cantonese line-up (CC, n = 23), Cantonese exposure and English line-up (CE, n = 22), English exposure and Cantonese line-up (EC, n = 23) and English exposure and English line-up (EE, n = 21). The experiment took place on an individual basis in a quiet room at the Chinese University of Hong Kong. It consisted of three stages: 1) the listeners listened to the target voice via headphones without being told what they would need to do later (exposure); 2) after exposure to the target voice, they took a five-minute mandatory break by filling in a questionnaire about their language background; 3) after the break, the participants were informed about the identification task. Further instructions were given and the line-up was played to the listeners on a computer. Two versions of the line-up in each language with different positions for the voices were used for counterbalancing: half of the participants listened to version 1 and half of the participants listened to version 2. The target voice was always present in the line-ups, but the listeners were not informed of this. The target voice was never positioned as the first or the last voice in the line-ups. The listeners listened to all seven 30-second speech samples one after another. At the end of the voice parade, they could repeat any of the samples as many times as they liked. Finally, they were asked to indicate on an answer sheet whether the target voice they were exposed to appeared in the parade and, if so, which one it was.[2] They were also asked to rate the confidence of their judgments on a 9-point scale (1 = no confidence at all, 9 = completely confident) and give the reason(s) for their judgment.

2.2 Results

2.2.1 Identification accuracy

Figure 1 shows the accuracy results of the Cantonese-English listeners in the four language conditions. Incorrect identification also included the 'target absent' answers, which were very few in number. In general, the listeners did not perform very well as incorrect answers outnumbered correct answers across all language conditions. Fisher's exact test confirmed that language conditions significantly affected listeners' accuracy (p = 0.041). The identification accuracy was better for the same-language conditions (EE and CC) than the cross-language conditions

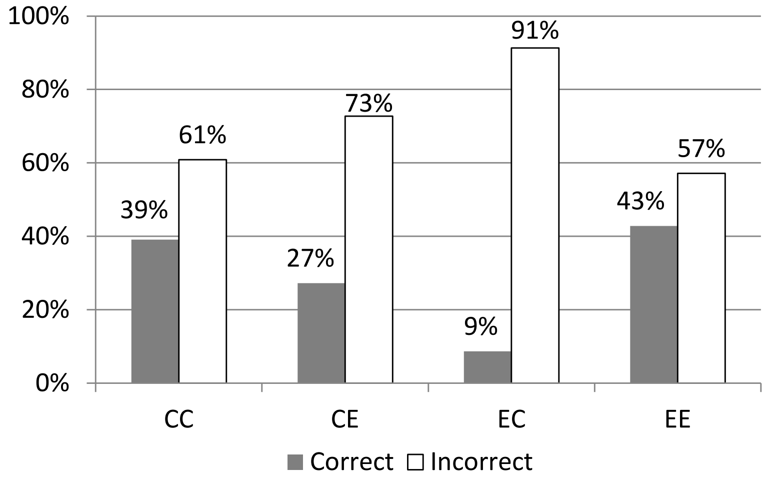

Figure 1: Percentage of correct and incorrect identification by Cantonese-English listeners in four language conditions

(EC and CE) (p = 0.021), while there was no significant difference between EE and CC, and between EC and CE (all p > 0.05). The chance level of choosing the target voice was 1/7 = 14.3%. It can be seen from Figure 1 that the accuracy of the EC condition was even below the chance level.

Two voices were auditorily quite similar to the target voice, particularly Foil A. They attracted many false alarms, and together they accounted for more errors than all the other foils put together, except in the EE condition. Table 2 shows the breakdown of incorrect identifications. The many false alarms for Foil A in the EC condition (52%) can explain why the accuracy of this condition was even below chance level.

Table 2: Breakdown (%) of incorrect identification by Cantonese-English listeners

	CC	CE	EC	EE
Foil A	35	27	52	10
Foil B	13	14	22	10
Others	13	32	17	37
All	61	73	91	57

2.2.2 Confidence rating

The listeners rated the confidence of their choices on a 9-point scale. The minimum chosen score was 3 and the maximum was 9. The mean confidence ratings of the four language conditions ranked as follows: CC (mean = 6.70, SD = 1.19) > CE (mean = 6.36, SD = 1.73) > EC (mean = 5.74, SD = 1.36) > EE (mean = 5.71, SD = 1.15). A one-way ANOVA indicated that the confidence ratings differed significantly across language conditions (F(3,85) = 2.769, p = 0.047). However, no significant pairwise comparison was found after applying the Bonferroni correction (all $p > 0.05$).

There was no correlation between accuracy and confidence (ρ = –0.004, p = 0.971). The confidence ratings for correct answers (mean = 6.12, SD = 1.37) and incorrect answers (mean = 6.14, SD = 1.44) were very similar. In addition, ten listeners with incorrect answers rated their confidence level to be 8 or 9. Conversely, only two listeners with correct answers rated their confidence level to be 8 or 9.

2.3 Discussion

The results clearly show that listeners performed better in the same-language than in the cross-language conditions, even though they had access to both linguistic and indexical information in all cases. As long as both the exposure and line-up were in the same language, whether it was in the listeners' native language or second language does not further affect listeners' identification accuracy, because linguistic and indexical information match in both CC and EE conditions. This corroborates the results in Köster and Schiller (1997), Winters et al. (2008) and Wester (2012) with bilingual data. On the contrary, the mismatch of linguistic and indexical information in the cross-language conditions resulted in a lower accuracy, which suggests that some indexical information may be language-dependent and is lost during the language switch.

Language conditions also affected listeners' confidence, but this time not in a simple same-language versus cross-language manner, as the listeners were most confident in the CC condition and least confident in the EE condition. It seems best to interpret the confidence data using the native versus non-native language perspective: they were most familiar with their native language, and thus were also most confident in it. The confidence data also echo those in previous studies in that there is no correlation between confidence and accuracy. Moreover, some listeners were very confidently wrong, and there were many more confident listeners with incorrect answers than those with correct answers. Listeners' confidence clearly cannot be used to assess the reliability of their identification.

3. Experiment 2: Mandarin-English

The Cantonese-English bilingual listeners in Experiment 1 had access to both linguistic and indexical information. Using a group of Mandarin-English bilingual listeners who did not understand Cantonese and exposing them to the same experimental material and procedure allows us to further evaluate the interaction between linguistic and indexical information in recognising voices. Based on the results of Experiment 1 and previous studies, it was hypothesized that Mandarin-English listeners would perform best in the EE condition, in which they have access to both linguistic and indexical information.

3.1 Method

3.1.1 Voices and materials

The same speech material as that employed in Experiment 1 was also used for Experiment 2.

3.1.2 Listeners

Eighty-nine Mandarin-English bilingual listeners participated in the identification experiment. They were all MA students at the Chinese University of Hong Kong, between 20 and 24 years old (mean = 22.46, SD = 0.72). They came from different regions of mainland China. Many of them also spoke a Chinese dialect in addition to Mandarin. They started learning English around the age of nine years and used English predominantly in academic settings. They participated in the experiment within the first three months of their arrival in Hong Kong. None of them could understand Cantonese. No history of hearing or speech problems was reported. All listeners received course credits for their participation.

3.1.3 Procedure

The voice ID procedure was identical to that employed in Experiment 1. The 89 Mandarin-English bilingual listeners were randomly assigned to one of the four language conditions: Cantonese exposure and Cantonese line-up (CC, n = 23), Cantonese exposure and English line-up (CE, n = 21), English exposure and Cantonese line-up (EC, n = 22) and English exposure and English line-up (EE, n = 23). Only in the EE condition did the listeners understand the speech material of the chosen voice in both the exposure and the line-up, and they did not understand anything in the CC condition.

3.2 Results

3.2.1 Identification accuracy

Figure 2 shows the identification results of the Mandarin-English listeners. The patterns are very similar across the four language conditions, with incorrect answers greatly outnumbering correct answers. Fisher's exact test confirmed that there was no significant difference across the four language conditions ($p = 0.973$). The accuracy in three language conditions (CE, EC, EE) was below chance level (14.3%), while that for CC was only slightly above chance level.

Akin to Experiment 1, the two voices that showed most similarity to the target voice had attracted many false alarms. Table 3 shows the breakdown of the listeners' incorrect responses. The two voices also accounted for more errors than the

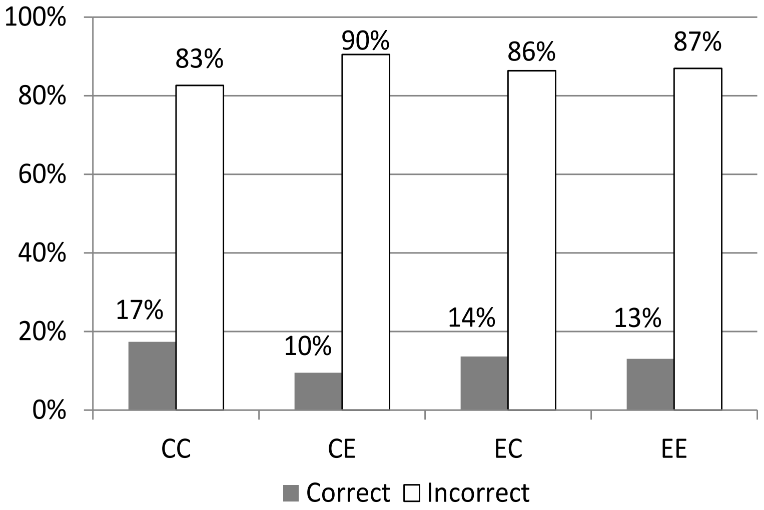

Figure 2: Percentage of correct and incorrect identification by Mandarin-English listeners in four language conditions

other voices together, but the differences between the two voices and other voices were not as large as those in Experiment 1.

Table 3: Breakdown (%) of incorrect identification by Mandarin-English listeners

	CC	CE	EC	EE
Foil A	35	24	14	52
Foil B	9	28	32	5
Others	39	38	40	30
All	83	90	86	87

3.2.2 Confidence rating

The listeners rated their confidence on a 9-point scale. The minimum chosen score was 1, and the maximum was 9. Their confidence ratings for the four language conditions ranked as follows: EE (mean = 6.57, SD = 1.50) > CE (mean = 5.86, SD = 2.06) > CC (mean = 5.61, SD = 1.70) > EC (mean = 5.05, SD = 1.81). One-way ANOVA showed that language conditions significantly affected listeners' confidence (F(3,85) = 2.853, p = 0.042). Post hoc tests with Bonferroni correction showed that the Mandarin-English listeners were significantly more confident in the EE than EC conditions (p = 0.03).

Similar to Experiment 1, there was no correlation between accuracy and confidence (ρ = −0.036, p = 0.734). The mean confidence rating for incorrect answers (5.81, SD = 1.84) was even higher than that for correct answers (5.58, SD = 1.78), although the difference was not significant. In addition, 14 listeners with incorrect answers rated their confidence level to be 8 or 9. Conversely, only one listener with a correct answer rated her confidence level to be 8.

3.3 Discussion

The identification accuracy results of the Mandarin-English listeners were unexpected. They performed equally poorly in all four language conditions, with accuracy only around chance level. Understanding the speech material or not did not affect their recognition accuracy. It was indeed surprising that no advantage was found in the EE condition. The results suggest that linguistic information and indexical information are processed independently of each other, different from the patterns in Experiment 1.

Unlike the accuracy results, Mandarin-English listeners' confidence ratings were affected by the language conditions. They were most confident in the EE condition, with scores even similar to the CC condition in Experiment 1. Again, their confidence level is not an accurate indication of their accuracy.

4. General Discussion

Identification accuracy results from the two experiments suggest that the relationship between linguistic and indexical information in voice recognition is not straightforward. A language effect (same-language vs cross-language) was found with the Cantonese-English listeners but no such effect was found with the Mandarin-English listeners. Cantonese-English listeners performed better than Mandarin-English listeners in general (compare Figures 1 and 2). Familiarity with the language(s) clearly influences how well listeners could remember a voice (Goggin et al. 1991; Winters et al. 2008). However, it is indeed intriguing to find that Mandarin-English listeners did not benefit from the EE condition in which they also had access to both linguistic and indexical information, while in the other conditions the linguistic information was either mismatched (CE and EC) or even absent (CC).

One likely reason for the lack of language effect for the Mandarin-English listeners is that they were unfamiliar with the speech features of Hong Kong English, which is an emergent new variety of English (Setter, Wong and Chan 2010). Pronunciation features of English spoken with a Cantonese accent are different from those of English spoken in China (Deterding 2006; Deterding, Wong and Kirkpatrick 2008). Familiarity with different accents can also affect voice recognition (Sjöström et al. 2008). Although the Mandarin-English listeners could understand the speech content spoken in Hong Kong English, they were unfamiliar with the specific phonetic/phonological features of the accent, including rhythmic and prosodic patterns. Therefore, they had only incomplete access to the linguistic information (see more discussion below). On a related note, another possibility is that some of the Mandarin listeners may not speak English as well as the Cantonese listeners given that they started learning English at a later age, and that they used English predominantly in school settings. Furthermore, many more Mandarin-English listeners were distracted by the similar sounding voice of Foil A in the EE conditions as compared to the other language conditions, and as compared to the Cantonese listeners (see Tables 2 and 3). These three reasons may explain the unexpectedly poor accuracy results of Mandarin-English listeners in the EE condition.

Although Cantonese-English listeners were familiar with both languages, they performed significantly better in the same-language than in the cross-language condition, echoing the findings for monolingual listeners (Wester 2012; Winters et al. 2008). The same-language advantage found for both monolingual and bilingual listeners suggests that it is not simply the presence or absence of linguistic and indexical information that affects their ability to recognise voices. It is likely that there are some language-dependent indexical cues to speaker identity which were lost with the language switch. Languages or even dialects can have different bases-of-articulation for 'phonemically equivalent' sounds (Bradlow 1995; Disner 1983; Jacewicz 1999, 2002; Recasens 2010; Torreira and Ernestus 2011). It should be noted that these differences are not linguistically contrastive or meaningful in these languages; they are just different 'settings'. Therefore, the same bilingual speaker can have distinct speech features in the two languages they speak.

In addition to the different articulatory settings of segmental features demonstrated by the studies cited above, bilingual speakers can also have distinct pitch 'settings' in their two languages. Stockmal et al. (2000) reported that their Korean-Japanese bilingual speaker with native-like proficiency in both languages spoke Japanese at a higher pitch than Korean. Such language-specific pitch difference is not idiosyncratic, as similar cross-language pitch differences were reported for Russian-English (Altenberg and Ferrand 2006), Mandarin-English (Xue, Hagstrom and Hao 2002) and Cantonese-English (Ng, Hsueh and Leung 2010) bilingual speakers. In fact, there is a growing literature showing that languages can differ in their pitch 'settings' (e.g., Keating and Guo 2012). Our data also show that most of the Cantonese-English bilingual speakers generally had a slightly higher pitch when speaking in English than when speaking in Cantonese (see Table 1). As pitch has been shown to be an important factor in voice recognition (Foulkes and Barron 2000; Sørensen 2012), the pitch difference, together with the potentially varying articulatory settings, could explain the poorer performance of the Cantonese-English listeners in the cross-language conditions. Therefore, although Winters et al. (2008) and Wester (2012) argued that there is sufficient language-independent indexical information in the speech signal for listeners to generalise knowledge of speakers' voices across languages, our data show that some indexical information can also be language-dependent. More studies on the systematic acoustic differences between the two languages of bilingual speakers are needed to further evaluate how the differences would affect the recognition of voices across languages.

If we compare our accuracy results with those in the literature, it seems that our bilingual listeners performed rather poorly. The best accuracy rate by the Cantonese-English listeners is only 42.9% (EE), while the accuracy rates of the Mandarin-English listeners just hover around chance level (14.3%). The retention interval (i.e., the time lag between exposure and identification) in our study was only five minutes, while in real life retention intervals are much longer. Higher recognition accuracy by monolingual listeners was reported in Sørensen (2012) (56–74%) and Thompson (1987) (38–65% in Experiment 1) with a retention interval of one week. Interestingly, accuracy rates comparable to our study were reported in Goggin et al. (1991) (12–57%, Experiments 1 and 2), who also used a five-minute retention interval.

Two reasons can be put forward that may explain the relatively poor performance of our listeners. Firstly, we intentionally used voices that are common and 'unmarked', and have included foil voices that were very similar to the target voice. These criteria are important in preparing a proper voice line-up (Broeders and van Amelsvoort 1999; Butcher 1996; Nolan 2003). As a result, the similar-sounding voices had attracted many false alarms (see Tables 2 and 3). This shows that the listeners could actually recognise some features of the target voice, but their memory was not accurate enough for them to identify the target voice correctly.

Secondly, and more importantly, we tested unexpected memory of the listeners by not informing them to remember the target voice during exposure to simulate more realistic forensic situations, while most other previous studies explicitly informed the listeners to remember the voice. Our listeners probably tried to understand and remember the content of the speech as much as possible during exposure. Therefore, their memory resources were mainly used to process and store the linguistic information rather than the indexical information. Informal inquiry after the experiment confirms this possibility. Many listeners reported that they thought that they would be tested on the content of the exposure speech, and were surprised by the identification task. Saslove and Yarmey (1980) demonstrated that informed listeners' voice memory was significantly better than that of uninformed listeners. The many studies reviewed by Yarmey (2007) also show that the identification accuracy of unfamiliar speakers heard only once is poor (less than 50%). Our poor bilingual accuracy results support Saslove and Yarmey's (1980) conclusion based on monolingual listeners that speaker identification is accurate only under the most favourable conditions and is extremely inaccurate in real-life forensic situations. Memory of unfamiliar voices is often unreliable, irrespective of language conditions.

It is interesting to see that, even with explicit instruction to remember the voice, a higher accuracy rate resulted with longer retention intervals, as demonstrated by Sørensen (2012) and Thompson (1987) (both with a one-week retention interval) versus Goggin et al. (1991) (five-minute retention interval). Of course, different methodological designs of these studies may be responsible for this contrast, but the same intriguing pattern was reported by Thompson (1987) with exactly the same experimental design: higher accuracy with a one-week interval (Experiment 1) than with a 25-minute interval (Experiment 3). Thompson suggested that the immediate interpolation of a task involving interaction with other voices between exposure and identification in Experiment 3 may have interfered with listeners' memory. However, it is unclear why many more activities with many voices in one week's time would not have the same negative impact on listeners' memory (Experiment 1). We do not have any plausible explanation for this intriguing contrast. The effect of time on memory decay is not simply proportional. In any case, it shows us that there are still many unknown factors influencing memory of voices which should be explored in future studies.

Language familiarity affects listeners' confidence in both experiments. Listeners were more confident in the same-language condition with their most familiar language (CC for Cantonese-English listeners, EE for Mandarin-English listeners). The same-language versus cross-language contrast does not seem to have any effect on listeners' confidence as the Cantonese-English listeners rated CC (mean = 6.70, SD = 1.19) to be one point higher than EE (mean = 5.71, SD = 1.15), with the two cross-language conditions in between. The listeners were more familiar with their own native language (Cantonese) than their second language (English), so they were more sensitive to the subtle features in the CC condition. Moreover, the emergent status of Hong Kong English means that there are variations in pronunciation (Sewell and Chan 2010) which may have increased listeners' uncertainty.

If we compare the confidence rating of the four language conditions in Experiment 1 and Experiment 2, an interesting pattern can be observed. Cantonese-English listeners (mean = 6.70) were significantly more confident than Mandarin-English listeners were (mean = 5.61) in the CC condition (t(44) = 2.517, p= 0.016), which is expected as the Cantonese-English listeners were listening to their native language while the Mandarin-English listeners did not understand anything in this condition. However, the Mandarin-English listeners (mean = 6.57) were significantly more confident than the Cantonese-English listeners were (mean = 5.71) in the EE condition (t(42) = −2.097, p = 0.042). Although the Mandarin-English listeners were not familiar with the pronunciation features of

Hong Kong English, they could understand the content only in the EE condition. The higher confidence level in this condition clearly shows the effect of language familiarity on listeners' confidence.

Despite the language effect, there is absolutely no correlation between confidence and accuracy, confirming similar findings in many previous studies (e.g., Goggin et al. 1991; Sørensen 2012; Thompson 1987; Yarmey 2001, 2004). Moreover, listeners can be completely confident and wrong. It is interesting to note that there were many more listeners having high confidence ratings but wrong answers than those with correct answers in our study. Orchard and Yarmey (1995) even found a significant negative confidence-accuracy correlation for distinctive voices, that is, the more confident the listeners were in the accuracy of their identification, the less correct they were in both target-present and target-absent line-ups. The confidence data clearly guard against the use of confidence as an indicator of identification reliability, contrary to the guidelines of the United States Supreme Court (*Neil v. Biggers* 1972).

In conclusion, our study illustrates the effects of bilingualism on speaker identification by lay witnesses in the same-language and cross-language conditions. There are complex interactions between linguistic information and indexical information. Some indexical information is language-dependent. In addition to the many factors on voice recognition identified by previous studies, the different 'articulatory settings' employed in different languages may further add to the difficulty of correct identification. Recognition of unfamiliar voices is often unreliable even with a very short retention interval. Thus, great caution must be taken when recognition of voice is needed, as in some forensic cases. Evidence gained from lay earwitness testimony must be treated with great caution. Indeed, in forensic casework this caution is strongly emphasised (IAFPA Code of Practice 6a and b), and the court is made aware of the limited reliability of lay witnesses to recognise voices – even more so in cross-language settings.

Acknowledgments

The authors would like to thank the Department of Linguistics and Modern Languages, The Chinese University of Hong Kong for supporting this research. They also thank the phonetics editors and two anonymous reviewers for helpful comments.

About the authors

Peggy Mok is an Associate Professor at the Department of Linguistics and Modern Languages, The Chinese University of Hong Kong. She received her MPhil and PhD in Linguistics from the University of Cambridge. Her research interests include experimental phonetics, psycholinguistics and bilingualism.

Robert Xu was an MPhil student at the Department of Linguistics and Modern Languages, The Chinese University of Hong Kong. His thesis was on cross-linguistic perception of intonation by Cantonese and Mandarin listeners. He is now studying for a PhD at Stanford University.

Donghui Zuo was an MPhil student at the Department of Linguistics and Modern Languages, The Chinese University of Hong Kong. Her thesis was on formant dynamics of diphthongs in the speech of bilingual identical twins, which is now published as a journal article.

Notes

1. The terms 'bilingual' and 'bilingualism' have many different meanings depending on the contexts in which they are used. In our study, we consider someone 'bilingual' if s/he knows a second language well enough to have an effective conversation in that language. All of our listeners have learnt English for over ten years.
2. There are several guidelines on how to prepare a voice line-up, e.g. the ones by Broeders and van Amelsvoort (1999) and the McFarland guidelines (see Nolan 2003 and the Home Office circular 057/2003 of the UK government). We followed the McFarland guidelines, which allow the listeners to make their decisions after they have heard all the samples in the line-up. The McFarland guidelines also allow the listeners to listen to any or all the samples as many times as they wish.

References

Abercrombie, D. (1967) *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Altenberg, E. P. and Ferrand, C. T. (2006) Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice* 20: 89–96. http://dx.doi.org/10.1016/j.jvoice.2005.01.005

Bradlow, A. R. (1995) A comparative acoustic study of English and Spanish vowels. *Journal of the Acoustical Society of America* 97(3): 1916–1924. http://dx.doi.org/10.1121/1.412064

Broeders, A. P. A. and van Amelsvoort, A. G. (1999) Lineup construction for forensic earwitness identification: a practical approach. Paper presented at the the 14th International Congress of Phonetic Sciences (ICPhS), San Francisco.

Butcher, A. (1996) Getting the voice line-up right: analysis of a multiple auditory confrontation. Paper presented at the the 6th Australian International Conference on Speech Science and Technology (SST), Adelaide.

Deterding, D. (2006) The pronunciation of English by speakers from China. *English World-Wide* 27: 175–198. http://dx.doi.org/10.1075/eww.27.2.04det

Deterding, D., Wong, J. and Kirkpatrick, A. (2008) The pronunciation of Hong Kong English. *English World-Wide* 29: 148–175. http://dx.doi.org/10.1075/eww.29.2.03det

Disner, S. (1983) Vowel quality. The relation between universal and language-specific factors. *UCLA Working Papers in Phonetics* 58.

Foulkes, P. and Barron, A. (2000) Telephone speaker recognition amongst members of a close social network. *International Journal of Speech, Language and the Law* 7(2): 180–198. http://dx.doi.org/10.1558/sll.2000.7.2.180

Goggin, J. P., Thompson, C. P., Strube, G. and Simental, L. R. (1991) The role of language familiarity in voice identification. *Memory and Cognition* 19(5): 448–458. http://dx.doi.org/10.3758/BF03199567

Goldstein, A. G., Knight, P., Bailis, K. and Conover, J. (1981) Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society* 17: 217–220. http://dx.doi.org/10.3758/BF03333718

Grosjean, F. (2013) Bilingualism: a short introduction. In F. Grosjean and P. Li (eds) *The Psycholingusitics of Bilingualism* 5–25. Hoboken: Wiley-Blackwell.

Hammersley, R. and Read, J. D. (1996) Voice identification by humans and computers. In S. L. Sporer, R. S. Malpass and G. Koehnken (eds) *Psychological Issues in Eyewitness Identification* 117–152. Mahwah, NJ: Lawrence Erlbaum.

Jacewicz, E. (1999). The base-of-articulation effect in a second language. Paper presented at the The 14th International Congress of Phonetic Sciences, Berkeley.

Jacewicz, E. (2002) The perception–production relationship in the acquisition of second language vowel contrasts. *Journal of Language and Linguistics* 1: 314–337.

Keating, P. and Guo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *Journal of the Acoustical Society of America* 132: 1050–1060. http://dx.doi.org/10.1121/1.4730893

Köster, O. and Schiller, N. O. (1997) Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics* 4(1): 18–28.

Köster, O., Schiller, N. O. and Künzel, H. (1995) The influence of native language background on speaker recognition. Paper presented at the the 13th International Congress of Phonetic Sciences (ICPhS), Stockholm.

Ng, M. L., Hsueh, G. and Leung, C. S. (2010) Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children. *International Journal of Speech-Language Pathology* 12(3): 230–236. http://dx.doi.org/10.3109/17549501003721080

Nolan, F. (2003) A recent voice parade. *International Journal of Speech, Language and the Law* 10: 277–291. http://dx.doi.org/10.1558/sll.2003.10.2.277

Orchard, T. L. and Yarmey, A. D. (1995) The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology* 9(3): 249–260. http://dx.doi.org/10.1002/acp.2350090306

Philippon, A. C., Cherryman, J., Bull, R. and Vrij, A. (2007) Earwitness identification performance: the effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology* 21: 539–550. http://dx.doi.org/10.1002/acp.1296

Recasens, D. (2010) Differences in base of articulation for consonants among Catalan dialects. *Phonetica* 67(4): 201–218. http://dx.doi.org/10.1159/000322312

Rogers, H. (1998) Foreign accent in voice discrimination: a case study. *Forensic Linguistics* 5(2): 203–208. http://dx.doi.org/10.1558/sll.1998.5.2.203

Saslove, H. and Yarmey, A. D. (1980) Long-term auditory memory: speaker identification. *Journal of Applied Psychology* 65(1): 111–116. http://dx.doi.org/10.1037/0021-9010.65.1.111

Schiller, N. O., Köster, O. and Duckworth, M. (1997) The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics* 4(1): 1–17.

Setter, J., Wong, C. S. P. and Chan, B. H. S. (2010) *Hong Kong English*. Edinburgh: Edinburgh University Press.

Sewell, A. and Chan, J. (2010) Patterns of variation in the consonantal phonology of Hong Kong English. *English World-Wide* 31(2): 138–161. http://dx.doi.org/10.1075/eww.31.2.02sew

Sjöström, M., Eriksson, E. J., Zetterholm, E. and Sullivan, K. P. H. (2008) A bidialectal experiment on voice identification. *Lund Working Papers in Linguistics* 53: 145–158.

Sørensen, M. H. (2012) Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech, Language and the Law* 19(2): 145–158. http://dx.doi.org/10.1558/ijsll.v19i2.145

Stockmal, V., Moates, D. R. and Bond, Z. S. (2000) Same talker, different language. *Applied Psycholinguistics* 21: 383–393. http://dx.doi.org/10.1017/S0142716400003052

Sullivan, K. P. H. and Schlichting, F. (2000) Speaker discrimination in a foreign language: first language environment, second language learners. *Forensic Linguistics* 7(1): 95–111. http://dx.doi.org/10.1558/sll.2000.7.1.95

Thompson, C. P. (1987) A language effect in voice identification. *Applied Cognitive Psychology* 1: 121–131. http://dx.doi.org/10.1002/acp.2350010205

Torreira, F. and Ernestus, M. (2011) Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology* 2(2): 331–353. http://dx.doi.org/10.1515/labphon.2011.012

Wester, M. (2012) Talker discrimination across languages. *Speech Communication* 54: 781–790. http://dx.doi.org/10.1016/j.specom.2012.01.006

Winters, S. J., Levi, S. V. and Pisoni, D. B. (2008) Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America* 123(6): 4524–4538. http://dx.doi.org/10.1121/1.2913046

Xue, A., Hagstrom, F. and Hao, G. (2002) Speaking fundamental frequency characteristics of bilingual Chinese-English speakers: a functional system approach. *Asia Pacific Journal of Speech, Language and Hearing* 7: 55–62. http://dx.doi.org/10.1179/136132802805576544

Yarmey, A. D. (1995) Earwitness speaker identification. *Psychology, Public Policy, and Law* 1(4): 792–816. http://dx.doi.org/10.1037/1076-8971.1.4.792

Yarmey, A. D. (2001) Earwitness descriptions and speaker identification. *Forensic Linguistics* 8(1): 113–122. http://dx.doi.org/10.1558/sll.2001.8.1.113

Yarmey, A. D. (2004) Common-sense beliefs, recognition and the identification of familiar and unfamiliar speakers from verbal and non-linguistic vocalizations. *International Journal of Speech, Language and the Law* 11(2): 267–277. http://dx.doi.org/10.1558/sll.2004.11.2.267

Yarmey, A. D. (2007) The psychology of speaker identification and earwitness memory. In R. C. L. Lindsay, D. F. Ross, J. Don Read and M. P. Toglia (eds) *The Handbook of Eyewitness Psychology* Vol. 2 *Memory for People* 101–136. Mahwah, NJ: Lawrence Erlbaum Associates.

Yarmey, A. D., Yarmey, A. L., Yarmey, M., J. and Parliament, L. (2001) Common sense beliefs and the identification of familiar voices. *Applied Cognitive Psychology* 15: 283–299. http://dx.doi.org/10.1002/acp.702

Case cited

Neil v. Biggers (1972) 409 U.S. 188, 199.