The Differential Effects of Subtitles on the Comprehension of Native English Connected Speech Varying in Types and Word Familiarity

Simpson W. L. Wong1, Cherry C. Y. Lin2, Isabella S. Y. Wong3, and Anisa Cheung1

Abstract
Connected speech produced by native speakers poses a challenge to second language learners. Video subtitles have been found to assist the decoding of English connected speech for learners of English as a foreign language (EFL). However, the presence of subtitles may divert the listeners’ attention to the visual cues while paying less attention to the speech signals. To test this proposal, we employed a bi-modal audio-visual listening test and examined whether EFL listeners were able to correctly identify the connected speech when misleading subtitles were present. We further tested whether connected speech with words of lower frequency further reduced the accuracy rate. Twenty-eight adolescent EFL learners, all with more than 10 years of experiences in learning English in schools, were tested with three major types of connected speech phonological processes, namely assimilation, elision, and juncture. The results of statistical analyses showed that matched and mismatched subtitles facilitated the comprehension of both familiar and unfamiliar connected speech. Error analyses revealed the degree of item-specific variations across the three types of connected speech processes as well as across the three subtitling conditions. This research provides insights on the immediate and long-term impact of subtitles on the decoding of English connected speech.

Keywords
listening, subtitles, connected speech, learning English as a foreign language, error analysis

Introduction
Learners of English as a foreign language (EFL) have been found to experience difficulties in comprehending English speech uttered by native speakers even after prolonged listening training in schools (Shockey, 2003). One reason is that most of the English speech (presented by non-native English-speaking teachers or pedagogically designed audios spoken by native speakers) appears in EFL classrooms is uttered at a slower rate to accommodate the EFL learners’ abilities. To ensure English words are properly learned, they are presented to EFL students in the citation form. Citation form, also known as dictionary form, is the way English words are presented in audio dictionary, that is, each word is presented in isolation and the pronunciation of the consonants and vowels are unaltered by surrounding phonological environment. According to Johnson (2004), approximately 60% of words in a corpus of 88,000 American English word tokens were spoken in reduced forms (i.e., speech signals are altered or reduced as compared with the utterance of a word in isolation). The more reduction and alternation of speech signal, the harder it is for EFL learners to perceive, as more demanding phonological reconstruction is required (Ernestus et al., 2002).

Of the numerous ways to improve listening comprehension, one of the most commonly adopted means is through the use of subtitles, which are frequently displayed in movies as closed captions and can be understood as a textual form of visual aid shown on screen. The usefulness of subtitles in video clips has attracted some scholarly attention in recent decades, and it appears researchers have reached the consensus that they can benefit different cohorts of learners,
including normal-hearing children and adults, as well as those with hearing impairments (Gernsbacher, 2015). In some empirical studies, subtitles are further classified as standard/inter-language (audio in a second language (L2) + subtitles in the first language (L1)), intra-language (audio in L2 + subtitles in L2), or reversed (dubbed audio in L1 + subtitles in L2). Through comparing performances between various subtitled conditions, studies on the effects of subtitling generally report a positive influence of intra-lingual (L2) subtitles in English (or other language) comprehension among second language learners (e.g., Bird & Williams, 2002). Similar findings of the positive effect of subtitles on English comprehension in animations, cartoons, movies, and TV series have been reported (Yang & Chang, 2014). Empirical studies on subtitling conducted so far mainly focused on the beneficial effects of subtitles on listening comprehension performance (e.g., Huang & Eskey, 2000), vocabulary acquisition, and recall of materials (e.g., Chun & Plass, 1996). Despite the presence of benefits provided by subtitles to EFL learners, subtitles are viewed differently by different EFL learners. Generally, lower-intermediate and intermediate learners do not have a choice by relying more on subtitles to comprehend the connected speech (Winke et al., 2010). It was revealed in an eye tracking study that the learners fixated on the subtitles 68% of the time. The interview data further suggested that these learners found reading subtitles easier than listening for extracting meaning and that reading also helped them more readily segment words from the stream of speech. These learners did not employ progressively fewer subtitles during the course of movie viewing (Pujola, 2002) and their learning goal was more oriented to improving reading rather than listening skills (Caimi, 2006; Chai & Erlam, 2008). In the long-run, without engagement with both the spoken and the subtitles signals, these less proficient EFL learners will hardly develop the ability to map phonological characteristics of sounds and the semantic meaning of words uttered in connected speech. Even worse, these groups of learners are less confident about their listening ability (Vanderplank, 2016).

Although general English language proficiency of the listeners plays a critical role in the use of subtitles (Yeldham, 2018), what awaits to be addressed is the large individual variations within the specific proficiency group. We speculate whether the nature of connected speech has an effect on listener’s performances. The two salient features about connected speech are the types of connected speech phonological processes as well as the familiarity of words within the connected speech. An introduction of the two features and relevant research findings is provided below.

**Processing Efficiency of Familiar and Unfamiliar Connected Speech**

As individual words are represented differently in connected speech, multiple phonological representations for individual words exist. A large body of research has shown a strong positive correlation between word frequency and the efficiency of auditory lexical access during spoken word recognition (e.g., Cleland et al., 2006; Connine et al., 1990, 1993; Goldinger, 1998; Marslen-Wilson, 1990), with the most frequently occurring words (high-production frequency) having the highest activation strength. In particular, the frequency at which the phonological variants are perceived in the context of continuous speech is shown to determine the efficiency of connected speech processing (Connine et al., 2008; Ernestus, 2014; Pitt & Samuel, 1995). For example, words that are frequently heard are recognized faster and more accurately in lexical decision tasks, as compared with those that are infrequently heard. Studies have also shown a reduced priming effect for low production frequency, compared with high production frequency, connected speech (Ranbom & Connine, 2007). Similarly, Mitterer and Russell (2013) demonstrated that higher production frequency is beneficial to the comprehension of reduced variants. Although the above studies examined the impact of word frequency on spoken word recognition in a uni-modal listening environment, it remains uncertain whether a similar word frequency effect can be observed in a bi-modal audio-visual listening context. The present study thus aims to fill this research gap.

**The Consistency of Performances Across Phonological Processes of Connected Speech**

For EFL learners, their first language (L1) background and L1 phonological system has been found to restrict the acquisition of L2 connected speech (Wong et al., 2017a). Among Chinese learners of English in particular, three connected speech phonological processes, namely assimilation, elision, and juncture, are largely affected by the cross-linguistic differences between Chinese and English sound systems. Assimilation is defined as the process in which phonemes assimilate to the place of the neighboring consonant while retaining their original voicing characteristics (Cruttenden, 2014). For example, the pronunciation of the word phrase /ten pletəz/ is changed from /ten 'pletəz/ to /tem 'pletəz/ in running speech. Assimilation is absent in Chinese utterances and therefore poses a challenge for Chinese EFL learners in English speech processing (Mao & Chen, 2013). In Liang’s (2015) study, 50 Chinese university sophomores majoring in English exhibited below-chance performances when identifying cases of assimilation. Another connected speech process that poses a challenge to Chinese EFL learners is (contextual) elision, in which a vowel or consonant occurring within either the body of a word or at a junction of word boundaries is lost (Cruttenden, 2014). This process is often evidenced in word phrases such as /iced tea/ (its citation form is /əst 'ti/), which is pronounced as /əst 'ti/ (similar to /ice tea/). The process of elision is also absent in Chinese utterances (Mao & Chen, 2013). In a study conducted in Taiwan, Chinese EFL sophomores with low, mid, and high English proficiency levels correctly identified only...
44%, 69%, and 77% of elision cases, respectively (Kuo, 2012). Finally, \textit{juncture} is the connected speech process referring to the removal of a clear boundary between two syllables. For example, /a name/ is transformed from its citation form [ə nɛm] to its reduced form [ə nɛm] (like /ən aim/). The obligatory gap between words to demarcate information is often absent in connected speech. In order to resolve juncture, a combination of contextual information and subtle cues in the speech signal is utilized by the listener to discern individual words (Setter et al., 2014). Moreover, it is suggested that English connected speech is like “singing in a legato way,” whereas Chinese connected speech is “articulated in a staccato way” (Duanmu, 2007, p. 71; Roach, 2008, p. 144). Junctures in the Chinese language can be clearly perceived due to the language’s syllable-timed pattern and the fact that the majority of words share the same degree of emphasis. As a result, Chinese EFL learners generally find it hard to identify junctures in English. When junctures are located before unstressed words or between words whose onsets or codas can be joined by the preceding or following words, they are almost unperceivable (Liang, 2015). Setter et al.’s (2014) empirical study indicated that Hong Kong listeners were only able to correctly identify 60% of junctures presented in British English speech. Given the suboptimal performances in Chinese EFL learners when identifying these three connected speech processes, it is therefore crucial to focus on these aspects in the present study.

**The Use of Subtitles in Non-immersive English Environments**

When learning English in a non-immersive English environment, EFL learners receive limited exposure to native English (Bradlow & Bent, 2002). Thus, these EFL learners have to rely heavily on listening materials in school settings as well as mass media, such as movies and TV channels (Vandergrift, 2011). As a common practice in Hong Kong, classroom materials are usually presented alongside a transcript, and the English-speaking movies are required to show Chinese subtitles (J. Y. H. Chan, 2013). Ironically, even if native English speech is readily available in non-English-speaking countries, listening skill is still not as good as expected (Wong et al., in press). This situation has motivated us to question the exact influences of subtitles on the decoding of connected speech.

In addition to the availability of listening materials, other factors such as language proficiency (e.g., Maleki & Rad, 2011), cognitive load of the multimedia (e.g., Winke et al., 2010), and design of subtitles (e.g., Chung, 1999) have also been investigated to understand the usefulness of subtitles. Still, a beneficial role of subtitles was assumed in these studies and the investigative focus remained on their degree of facilitation. However, as implied by the poor listening comprehension performances in EFL learners with various language backgrounds who can readily access subtitles (e.g., Chinese: Chung, 1999; French: Guillory, 1998; Spanish: Markham & Peter, 2003; Russian: Winke et al., 2010), we suspect that subtitles may not facilitate the acquisition of connected speech under all circumstances and reducing any anxieties triggered when listening to non-L1 speech (Behroozizzad & Majidi, 2015), the negative aspects of subtitles require further investigation.

In an audio-visual environment, it is well known that the visual modality is dominant, as visual information is processed significantly faster than auditory information; hence, auditory processing is often subduced (Lukas et al., 2010). Moreover, if subtitles are treated as \textit{obligatory} rather than \textit{supplementary} during listening, EFL learners are likely to experience listening difficulties when no subtitles are provided (Gruguović & Hegelheimer, 2007; Pujola, 2002). Due to the more enduring nature on the screen of the subtitles than the connected speech (Hulstijn, 2003), ESL learners may develop the tendency to read subtitles for extracting meaning and/or segmenting words in the stream of connected speech (Winke et al., 2010). One negative effect we speculate is “subtitle-dependency,” whereby learners are more likely to trust what they read from subtitles over what they hear from the corresponding audio when they are presented with mismatched subtitles, that is, when the content of the subtitles does not match the content of the audio. Hence, we predict that EFL learners are more prone to listening errors when mismatched subtitles are provided.

**The Present Study**

In this section, we shift our focus and provide a brief description of the context of the present study. In Hong Kong, English language is a compulsory subject and is introduced in the curriculum as early as 3 years of age. Despite the official status of the language, it is not commonly spoken in daily life conversation. Students are, however, highly motivated to get good results in English, as proficiency in English directly affects the results of public examination and thus their chance of university admission. Apart from academic use, students in Hong Kong are also exposed to spoken native English through various forms of media such as movies and songs, as well as having lessons with their Native English-speaking teachers in school.

Despite the rich environment available for English learning, it is surprising to note that Hong Kong undergraduates who have been learning English for over 15 years were still unable to fully decode connected speech spoken by native English speakers (Wong et al., 2017b); furthermore, in an earlier study, Shockey (2003) found that Hong Kong Cantonese speakers who were English language teacher trainees made multiple perceptual errors when decoding English connected speech. It is essential to understand whether high school EFL students in Hong Kong are dependent on subtitles during listening comprehension. Although this enhances immediate decoding of speech, the use of subtitles may hinder auditory
perceptual learning of connected speech, causing long-term suboptimal performances in connected speech processing. This potential negative effect motivates us to study how connected speech is processed when subtitles are presented. Do listeners trust their ears or their eyes more? In addition, we compare listeners’ accuracy in decoding speech across three types of connected speech (assimilation, elision and juncture) in order to reveal their unique characteristics and the level of difficulties they represent to listeners.

In light of the aforementioned research gaps, the four research questions of the present study are listed as follows:

**Research Question 1:** How do EFL learners perform in decoding connected speech with and without the presence of subtitles?

**Research Question 2:** Do matched subtitles facilitate, and mismatched subtitles conversely interfere with, connected speech decoding?

**Research Question 3:** Does the familiarity of the words affect decoding performance?

**Research Question 4:** Are performance levels consistent across the three types of connected speech processes (assimilation, elision, and juncture)?

**Method**

**Participants**

A total of 28 Cantonese-speaking EFL learners (16 females; 12 males) aged between 15 and 16 years were recruited for the current study. The students had no reported difficulties in learning or language acquisition. All participants were 10th graders from a local secondary school in which Cantonese is used as the medium-of-instruction for non-English subjects. These students were recruited by the second author, who was their English teacher. Based on the benchmark against the territory-wide English standard specified by the Hong Kong Education Bureau (2004) as well as via a medium-level listening test (Davis, 1998), the participants were rated “average” in terms of English proficiency.

**Procedure**

This work was conducted with the formal approval of Human Research Ethics Committee at (The Education University of Hong Kong). After obtaining informed consent from participants and their parents, respectively, the experiment was carried out in a classroom in the participants’ school. All students were presented with the same set of testing materials in a group-testing situation. The uni-modal connected speech decoding task (presented to the participants as an unseen dictation task) was administered first, followed by the bi-modal connected speech decoding test. The test administration order was designed in such a way to avoid the influence of prior exposure to the subtitles in the bi-modal task on the uni-modal task.

**Preparation of Audio**

We first generated a list of minimal pairs of connected speech while taking the English proficiency levels of the participants into consideration. The list was developed by our research team and was only adopted in the current study. The stimuli were recorded in a soundproof room using a high-quality Roland R-09HR recorder and digitized at a sample rate of 44.1 kHz with a 16-bit amplitude resolution. The set of stimuli was spoken aloud by a 25-year-old native female speaker who was born and raised in New York. This speaker also advised on the suitability of the test items for the use in this study. A General American (GA) accent was chosen for this study as Hollywood movies and TV programs are popular among Hong Kong Chinese adolescents and they are commonly exposed to this accent. Furthermore, as shown in J. Y. H. Chan’s (2013) study, young adults in Hong Kong regard the GA accent as native accent.

**Measures**

**Bi-modal audio-visual connected speech comprehension test.** This test assesses listeners’ ability to decode minimal pairs of connected speech, distinguished by one differing segment within the whole phrase. By manipulating two parameters, namely the link between the content of the spoken word phrases and the subtitles (matched vs. mismatched) and the degree of word phrase frequency or “familiarity” (familiar vs. unfamiliar), four listening conditions were created: (a) a familiar phrase with matched subtitles, (b) a familiar phrase with mismatched subtitles, (c) an unfamiliar phrase with matched subtitles, and (d) an unfamiliar phrase with mismatched subtitles. The familiar items were connected speech phrases comprising high-frequency words or formulaic phrases (e.g., not at all). Conversely, the resulting unfamiliar items in the minimal pairs contained unfamiliar lexical items and/or non-formulaic phrases. It should be noted that “unfamiliar” phrases such as /hot takes/ may not make sense under normal circumstances. Yet, we cannot rule out the possibility that the phrase may be articulated in a playful situation. Thus, these phrases are regarded as unfamiliar instead of illegitimate.

The recordings were deliberately designed to include connected speech processes that have been shown to be difficult for Chinese EFL learners, namely assimilation, elision, and juncture (A. Y. Chan & Li, 2000; Liang, 2015; Setter et al., 2014; Yang & Chang, 2014). The stimuli were short phrases without filler sentences. As shown in Table 1, the number of words in canonical/citation form was limited to 5 to minimize the working memory load. They also did not carry contextual cues for top-down, meaning-driven predictions. The same set of stimuli was presented in the subtitles-matched and subtitles-mismatched conditions. The number of words was therefore the same across the two conditions. There were altogether 36 minimal pairs of target connected speech patterns (see Table 1 for details). As
the whole battery of test can be potentially be used for language assessment, scale reliability was examined. We computed Cronbach’s α that inform the degree of relatedness among a set of test items in a battery of tests. Reliability of the whole test (k = 72) was .73 (Cronbach’s α) which was commonly recognized as acceptable to good reliability (Cortina, 1993).

Participants were told that the recordings would be played with either matched or mismatched subtitles. They were therefore encouraged to focus on aural, as opposed to visual, information when answering the multiple-choice questions. For each trial, the student participants were presented with audio and subtitles concurrently through a speaker and PowerPoint slides (Figure 1). As the experiment was run in a group setting, the experimenters had to make sure that the participants completed each item before proceeding to the next item. Therefore, the bi-modal listening test was controlled and paced by the experimenter.
For half of the trials, the audio stimuli did not match the content of the subtitles, that is, they were presented in the “mismatched” condition. For example, the audio of [I dome know] was played together with the subtitle /I don’t know/. The remaining half of the trials presented matched audio and subtitles, that is, the “matched” condition. For example, /I don’t know/ was presented in both auditory and visual forms (Figure 2). Students were given two choices on the answer sheet, for this example the choices would be “I dome know” and ‘I don’t know.’ They were instructed to listen and select the phrase that they heard from the two choices printed on the test paper. These 72 items were randomized for presentation. On the randomization list, 64% of the items were presented with the subtitles-matched condition first and 36% of them were presented with the subtitles-mismatched condition first.

In addition, the familiar and unfamiliar items in the phrase pairs were presented first in 47% and 53% of the items. These percentages could minimize the possibility that listeners’ decision was conditioned by their decision of the previous items. Each correct response yields one mark, and the number of correct responses in each of the listening conditions was summed to yield a mean score for further statistical analyses.

**Uni-modal connected speech comprehension test.** In the “no-subtitles” condition, the experimenter played the audio stimuli in a fixed order and asked listeners to dictate what they heard. No subtitles were provided within this condition. The total number of test items was 36. An all-or-none scoring criterion was adopted such that an answer is considered correct only if it matches exactly the same as our answer key (Table 1). The marking was carried out by the first and second authors and a 100% agreement was achieved. Each correct response yields one mark, and summing the correct responses would produce a mean score for comparing the listening performances across conditions.

**Results**

A mix of inferential and descriptive statistical analyses were performed to investigate the influence of subtitles on decoding three types of English connected speech in Chinese EFL learners. The data analyses were conducted on data obtained from the 28 participants who finished all the tests. With a 3 (type of connected speech: assimilation vs. elision vs. juncture) × 2 (familiarity: familiar vs. unfamiliar) × 3 (subtitling: without subtitles vs. with matched subtitles vs. with mismatched subtitles) design, yielding a total of eighteen listening conditions. The means and standard deviation of these conditions are listed in Table 2.

Repeated-measures analysis of variance (ANOVA) with two within-subject factors was used to compare the performances across various listening conditions. Before the analysis, we tested the assumption using Mauchly’s test of sphericity. If this assumption is violated ($p < .05$), a correction of degrees of freedom by Huynh–Feldt estimates of sphericity was carried out (Field, 2013, p. 474). When a significant interaction was detected, contrast analysis was conducted to examine where the difference lied.

**Connected Speech Perception Performance in the No-Subtitles Condition**

First, we assessed the performance of connected speech decoding in the “no-subtitles” condition, by comparing the three types of connected speech and examined the effect of familiarity. A 3 (types: assimilation vs. elision vs. juncture) × 2 (familiarity: familiar vs. unfamiliar) repeated-measures ANOVA was computed. The main effect of types was significant, $F(2, 54) = 29.96, p < .001, \eta^2_p = .52$. A post hoc test of least significant
difference (LSD) indicated that the performance of decoding assimilation was significantly better than decoding elision ($p = .003$) and juncture ($p < .001$). Moreover, the performance of decoding elision was significantly better than decoding juncture ($p < .001$). The main effect of familiarity was significant, $F(1, 27) = 112.52, p < .001, \eta_p^2 = .80$, indicating that familiar items were better decoded than unfamiliar ones. The interaction between types and familiarity was significant, $F(2, 54) = 33.41, p < .001, \eta_p^2 = .55$. Contrast analysis showed that assimilation was more sensitive to familiarity than elision, $F(1, 27) = 28.74, p < .001, \eta_p^2 = .51$, and juncture, $F(1, 27) = 92.74, p < .001, \eta_p^2 = .77$, thus indicating that the change (improvement) of performances from unfamiliar to familiar items was larger in assimilation than in elision and juncture.

The Influence of Subtitles on Decoding Assimilation, Elision, and Juncture

Next, we examined the interplay between subtitling and familiarity. We conducted a 3 (subtitling: without subtitles vs. with matched subtitles vs. with mismatched subtitles) $\times$ 2 (familiarity: familiar vs. unfamiliar) repeated-measures ANOVA for each of the three types of connected speech, namely assimilation, elision, and juncture. The results will allow us to evaluate whether the two hypothesized variables have the same effect on different types of connected speech.

Assimilation. The main effect of subtitling was significant, $F(2, 54) = 301.717, p < .001, \eta_p^2 = .91$. Post hoc comparison by LSD revealed that the scores obtained in the matched-subtitles condition were significantly higher than those obtained in the mismatched-subtitles ($p = .001$) and no-subtitles conditions ($p < .001$). In addition, the scores obtained in the mismatched-subtitles condition were significantly higher than those obtained in the no-subtitles condition ($p < .001$). The main effect of familiarity was significant, $F(1, 27) = 126.00, p < .001, \eta_p^2 = .82$, indicating that familiar items were better decoded than unfamiliar items. The subtitling $\times$ familiarity interaction was significant, $F(2, 54) = 57.95, p < .001, \eta_p^2 = .68$. Contrast analysis revealed that performances in the no-subtitles condition were more sensitive to familiarity than the matched-subtitles condition, $F(1, 27) = 64.72, p < .001, \eta_p^2 = .70$, and mismatched-subtitles condition, $F(1, 27) = 87.57, p < .001, \eta_p^2 = .76$. In other words, the change (improvement) of decoding performance from unfamiliar to familiar items was larger in the no-subtitles condition than in the matched- and mismatched-subtitles conditions.

Elision. Mauchly’s test indicated that the assumption of sphericity for the main effect of subtitling had been violated, $\chi^2(2) = 17.06, p < .001$; therefore, degrees of freedom were corrected using Huynh-Feldt estimate of sphericity ($e = .69$). The main effect of subtitling was significant, $F(1.13, 36.45) = 455.55, p < .001, \eta_p^2 = .94$. Post hoc comparison by LSD revealed that the scores obtained in the matched- and mismatched-subtitles condition were significantly higher than those obtained in the no-subtitles condition ($ps < .001$). In addition, the scores obtained in the matched-subtitles and mismatched-subtitles condition were not significantly different ($p = .83$). On the other hand, the main effect of familiarity was significant, $F(1, 27) = 12.88, p = .001, \eta_p^2 = .32$, indicating that familiar items were better decoded than unfamiliar items. The subtitling $\times$ familiarity interaction was significant, $F(2, 54) = 23.99, p < .001, \eta_p^2 = .47$. Contrast analysis revealed that performances in the no-subtitles condition were more sensitive to familiarity than those in the mismatched-subtitles condition, $F(1, 27) = 22.71, p < .001, \eta_p^2 = .45$, and mismatched-subtitles condition, $F(1, 27) = 39.87, p < .001, \eta_p^2 = .59$. In other words, the change (improvement) of decoding performance from unfamiliar to familiar items was larger in the no-subtitles condition than in the matched- and mismatched-subtitles conditions.

Juncture. Mauchly’s test indicated that the assumption of sphericity for the main effect of subtitling and the interaction effect between subtitling and familiarity had been violated, $\chi^2(2) = 6.52, p = .038$ and $\chi^2(2) = 11.29, p = .004$, respectively. Therefore, degrees of freedom were corrected using Huynh-Feldt estimate of sphericity ($e = .86$ and $e = .77$). The main effect of subtitling was significant, $F(1.72, 46.65) = 780.91, p < .001, \eta_p^2 = .96$. Post hoc comparison by LSD revealed that the scores obtained in the matched- and mismatched-subtitles conditions were significantly higher than
In summary, decoding of speech embedded with assimilation, elision, and juncture was poor when no subtitles were provided. Interestingly, the matched subtitles did not always serve as better subtitles than mismatched subtitles as a significant difference was found only among the assimilation items. Accuracy in decoding familiar connected speech was higher than unfamiliar connected speech in assimilation and elision only.

**Item Analysis Within Each of the Three Types of Connected Speech**

Subsequently, we performed item analysis to further compare listeners’ decoding performances across the three subtitling conditions for each individual test item. Specifically, we examined whether speech tokens under the same category (e.g., assimilation) elicited similar levels of accuracy. As
shown in Tables 3 to 5, the range of percentage accuracy in speech identification was large across the three types of connected speech as well as across the three subtitling conditions, suggesting large item-specific variations.

As noted in the no-subtitles condition, the maximum percentage accuracy in speech identification for assimilation was 100% and the minimum was 0%, yielding a range of 100%. The ranges for elision and juncture in the same condition were 64% and 57.1%, respectively. In both matched-subtitles and mismatched-subtitles conditions, a 100% accuracy rate could be obtained for some items across the three types of connected speech. Although the ranges of percentage correct were similar for elision and juncture (about 30%), the range for assimilation was almost double that of the former two types (about 60%). As shown in Table 3, a number of inaccurate regressive assimilation errors was recorded, particularly for unfamiliar items. For instance, participants commonly misinterpret “tank coins” /ˈtæŋkˌkɔɪnz/ as “ten coins” /ˈtæŋˌkɔɪnz/. For speech items embedded with juncture, a constant pattern of mis-segmentation was observed, such as misinterpreting “not at all” as “not that tall.”

In general, the performances of decoding assimilation, elision, and juncture were enhanced with the presence of either matched or mismatched subtitles. However, it is worth noting one exception. The accuracy rate of perceiving “I don’t know” is in fact lower in the conditions with subtitles (82.1%) compared with the no-subtitles condition (100%). Based on the above results, it is suggested that individual items vary in terms of level of difficulties as well as the sensitivity to subtitles.

**Discussion**

The present study aims to examine the role of subtitles in connected speech decoding in Chinese EFL learners. We used both matched and mismatched subtitles to test how listeners process visual and audio information in the face of conflicting situations. Our results showed that the performances of decoding connected speech in both matched and mismatched subtitles could significantly facilitate the processing of the three types of connected speech examined, which is generally in line with previous studies about the usefulness of matched subtitles in listening. These findings can strengthen the claim that provision of subtitles is crucial for decoding native English connected speech for Chinese EFL learners. More importantly, the data in the mismatched condition provide new insight to the use of subtitles in listening comprehension among EFL learners, suggesting that EFL adolescent learners who have not yet attained native-like English listening skills still attempted to link visual information with audio information, rather than disregard the latter completely in a bi-modal listening environment. Given that the experimental stimuli used in the present study were phrases of connected speech that differed only by one segment, the participants demonstrated their abilities to identify the critical segments that distinguished the minimal pairs of connected speech.

The findings of significant facilitation by mismatched subtitles further suggest that the potential detrimental effect of visual-dominance is minimal in a bi-modal listening environment (Lukas et al., 2010). Listeners were found to be able to process the auditory signals even in the presence of misleading visual information. Furthermore, our results suggest that the listeners had sufficient executive functioning to inhibit irrelevant information as well as to selectively attend to audio information (Leon-Carrion et al., 2004). However, it is important to note that our participants were only exposed to subtitles and audio. Taking the limited capacity of cognitive resources into consideration (Drijvers et al., 2016), a trade-off between auditory and visual information processing occurs in a bi-modal listening environment. If video is present as in movies and TV programs, listeners’ cognitive

| Familiarity | Test items       | Without subtitles | With matched subtitles | With mismatched subtitles |
|------------|------------------|-------------------|------------------------|---------------------------|
| F          | Not at all       | 0                 | 75.0                   | 71.4                      |
| U          | Not that tall    | 3.6               | 75.0                   | 71.4                      |
| F          | It wasn’t easy  | 39.3              | 92.9                   | 92.9                      |
| U          | It was easy     | 0                 | 94.9                   | 82.1                      |
| F          | Better off      | 3.6               | 96.4                   | 92.9                      |
| U          | Better ruff     | 0                 | 100                    | 92.9                      |
| F          | Leave it to me  | 7.1               | 85.7                   | 78.6                      |
| U          | Lean fit to me  | 3.6               | 89.3                   | 82.1                      |
| F          | No apple        | 57.1              | 85.7                   | 100                       |
| U          | An old wapple   | 10.7              | 92.9                   | 85.7                      |
| F          | One wish each   | 0                 | 92.9                   | 96.4                      |
| U          | One wish sheet  | 0                 | 100                    | 89.3                      |

F = familiar; U = unfamiliar.
load may be further increased, leaving little attentional resources to process audio information. Therefore, the use of multimedia in training connected speech decoding needs to take into account the nature and amount of visual information presented to the listeners.

As evidenced across the three subtitling conditions, familiar word phrases were better recognized than unfamiliar word phrases. This implies that novel connected speech processes and signals are challenging to EFL listeners, whose phonological repertoire for decoding connected speech may not be versatile enough. We speculate that the immediate success of decoding connected speech with the aid of subtitles may hinder long-term perceptual training. As previously described by Schnitz and Kurschner (2007), cognitive task performance and learning are different. Cognitive task performances are those actions that operate on mental structures in working memory, whereas learning operates on mental structures in long-term memory. In other words, learning takes place only if the content in long-term memory is transformed and results in an increase in expertise. If EFL learners are used to processing connected speech in working memory for immediate outcome without activating the phonological repertoire in long-term memory, their listening skills cannot be improved even after constant exposure to native English speech. As a result, when encountering novel connected speech without subtitles, subtitle-dependent EFL listeners are more likely to struggle in decoding the connected speech signal, though this claim has yet to be verified in future studies. The ingrained “performing without learning” situation is further reinforced by the use of (or reliance on) subtitles. Thus, it is important to prevent the development of this vicious circle within connected speech acquisition.

The Influence of Frequency on Connected Speech Perception

The current study examines perception of speech embedded with three types of connected speech processes (assimilation, elision, and juncture). When considering word frequency within the no-subtitles condition, the identification of high familiarity words was higher than those of low familiarity. In other words, these three connected speech processes are frequency-sensitive: The higher the frequency, the higher the accuracy, and vice versa. However, when subtitling was considered, the three types of connected speech processes display varying degrees of frequency-sensitivity when presented without subtitles. Yet, both assimilation and elision were found to be more sensitive to familiarity effects than juncture. One possible explanation for this finding is the perceptual differences between various connected speech processes. For assimilation and elision, it may be that the neighboring phonemes serve as cues which contribute to the identification of modified segments. This hypothesis is consistent with the view of probabilistic phonotactics (see Vitevitch & Luce, 1999). When viewed in conjunction with the theory that these processes are segmental “overlaps” with gradient variations (e.g., Browman & Goldstein, 1992), residual speech sounds between neighboring phonemes’ transitions may be perceived. If this is the case, the frequency effect will become more beneficial as listeners become more experienced. This is due to the shorter phonetic distance and less effortful lexical activation that this entails. In contrast, for juncture, word boundaries are ambiguous. For this process, non-native listeners tend to rely more on sentential and lexical context to aid segmentation as opposed to acoustic-phonetic clues (Altenberg, 2005; Mattys & Melhorn, 2007). As only non-contextual minimal pairs were used in the study, the stimuli lacked supporting top-down information. Therefore, participants relied on limited speech decoding abilities for word segmentation. For these cases, frequency would impose a minimal effect on juncture, as acoustic-phonetic cues may not be the major route utilized for comprehension. Furthermore, the phonetic distance may not be as short as that exhibited for assimilation and elision. These assumptions call for further research to investigate whether EFL learners adopt different speech-decoding strategies for different connected speech processes.

Last but not least, it should be noted that listening performances were not perfect even when accurate subtitles were provided. It is possible that some EFL learners may not be able to decode the subtitles because of inadequate reading skill. Although we did not measure reading ability in the current study, our speculation is partly supported by several previous studies showing a strong link between overall language proficiency and the efficiency of subtitle use during listening tasks in EFL learners (e.g., Lwo & Lin, 2012; Maleki & Rad, 2011). Still, further research is needed to verify the specific links between reading skills and connected speech processing skills.

Pedagogical Implications

In line with other studies examining connected speech in EFL learners, suboptimal connected speech comprehension skills were observed in the current sample. This result again pinpoints the need for connected speech listening training in EFL classrooms. As implied in the current study, using onscreen subtitles for educational purposes should be promoted but with greater caution. Pedagogically speaking, the ideal situation is that EFL learners continuously attempt to connect the content of the subtitles and audio when being exposed to them. This is key to ensuring that characteristics of English speech are learned, such as phonotactic properties, novel vocabulary, and grammar of English sounds (Brown, 2012). However, multimedia may only be used for entertainment purposes and the motivation to learn connected speech may not be consistently high across EFL learners. In reality, EFL learners may devote most of their attention to the subtitles and very little attention would be allocated to the audio input. Teachers are advised to remind
students of both the positive and negative effects of subtitles on listening training. The focus of English listening comprehension should be placed on training students to comprehend native English connected speech without relying on subtitles. Thus, upon the use of subtitles (or other visual cues) to facilitate the learning of English lexical knowledge, teachers should gradually remove the visual cues so that students can comprehend English connected speech eventually in a subtitles-free listening environment. In addition, the significant familiarity effect obtained in the present study suggests that there is a need to consider this linguistic parameter when planning the listening curriculum. Teachers may design a variety of learning activities to consolidate the lexical familiarity of their students, so as to facilitate listening comprehension. It is interesting to note that assimilation was more sensitive to familiarity than elision. Teachers can tell students this characteristic explicitly and should be more alert to the familiarity effect when training different types of connected speech. Teachers should also choose captioned videos that are of appropriate level to the students.

However, the familiarity effect should not be over-interpreted as having to pre-teach all difficult vocabulary/phrases from a text prior to the introduction of a listening task (see Liao & Yeldham, 2015). Even though this is believed to scaffold their comprehension and to heighten their awareness of new items in the text (e.g., Chai & Erlam, 2008), doing so removes the chance for learners to practice inferring the meaning of such items from the context. Hence, teachers should strike a balance between giving vocabulary input and nourishing students’ skills on drawing inferences from the speech.

Limitations and Future Research Directions

The first limitation concerns the number of items for each of the listening conditions. In order to create minimal pairs for the selected connected speech processes, the script was written with a focus on phrases with similar pronunciation. Such constraint, together with another parameter we manipulated (i.e., word familiarity) had limited the number of minimal pairs we could generate. Although the number of items could be increased, the effect size obtained from this study was large enough to substantiate our results. In future studies, the inclusion of additional items is recommended.

Another limitation is the presentation of subtitles during the test. The participants were provided with a list of minimal pairs and were requested to listen and choose to make the test procedure easier to understand. However, the option of force-choice questions may provide additional visual cues for participants as they were not required to provide the answers themselves. Given the provision of differential visual cues and the difference in response format, a direct comparison between the performances under the subtitle conditions and no-subtitles condition is deemed to be unfair.

Future studies could therefore include a dictation test in the subtitled conditions.

Conclusion

The usefulness of subtitles for the immediate decoding of connected speech in a bi-modal listening setting is well documented in the literature. However, it is worth reconsidering the impact subtitles have on the acquisition of acoustic, phonetic, and phonological aspects of connected speech in the long run. Consistent with previous research on subtitles, subtitles are shown to be beneficial to listening comprehension in the current study. Hence, we do not advocate to abandon them, especially as EFL learners were found to favor the use of subtitles in their daily lives (Dallas et al., 2016). Still, EFL learners are reminded to equip themselves to become visual-aid-free competent listeners. Moreover, instructors or self-taught learners should be mindful about the exact function of subtitles for various listeners and be able to shift their focus back and forth between audio and visual information.

Acknowledgment

We would like to thank all the participants.

Data Availability

The data that support the findings of this study are available from the corresponding author (S.W.L.W.), upon reasonable request.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by Early Career Scheme of the Research Grants Council (RGC) of Hong Kong (ECS 846212).

ORCID iD

Simpson W. L. Wong https://orcid.org/0000-0002-6606-6382

References

Altenberg, E. P. (2005). The perception of word boundaries in a second language. *Second Language Research, 21*(4), 325–358. https://doi.org/10.1191/0267658305sr250oa

Behroozizad, S., & Majidi, S. (2015). The effect of different modes of English captioning on EFL learners’ general listening comprehension: Full text vs. keyword captions. *Advances in Language & Literary Studies, 6*(4), 115–121. https://doi.org/10.7575/aiac.allsv.6n.4p.115

Bird, S. A., & Williams, J. N. (2002). The effect of bimodal input on implicit and explicit memory: An investigation into the ben-
effects of within-language subtitling. *Applied Psycholinguistics*, 23(4), 509–533. https://doi.org/10.1017/S0142716402004022

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112*(1), 272–284. https://doi.org/10.1121/1.1487837

Brown, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, 49(3–4), 155–180. https://doi.org/10.1159/000261913

Brown, J. D. (Ed.). (2012). *New ways in teaching connected speech*. Teachers of English to Speakers of Other Languages.

Cain, A. (2006). Audiovisual translation and language learning: The promotion of intralingual subtitles. *The Journal of Specialised Translation, 6*, 85–98.

Chai, J., & Erlam, R. (2008). The effect and the influence of the use of video and captions on second language learning. *New Zealand Applied Linguistics, 14*, 25–44.

Chan, A. Y., & Li, D. C. (2000). English and Cantonese phonology in contrast: Explaining Cantonese ESL learners’ English pronunciation problems. *Language Culture and Curriculum, 13*(1), 67–85. https://doi.org/10.1080/07908310008666650

Chan, J. Y. H. (2013). Contextual variation and Hong Kong English. *World Englishes, 32*(1), 54–74. https://doi.org/10.1111/weng.12004

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotation on vocabulary acquisition. *The Modern Language Journal, 80*(2), 183–198. https://doi.org/10.2307/328635

Chung, J. (1999). The effects of using video texts supported with advance organizers and captions on Chinese college students’ listening comprehension: An empirical study. *Foreign Language Annals, 32*(3), 295–308. http://dx.doi.org/10.18806/tels.v14r1.678

Cleland, A. A., Gaskell, M. G., Quinlan, P. T., & Tamminen, J. (2006). Frequency effects in spoken and visual word recognition: Evidence from dual-task methodologies. *Journal of Experimental Psychology: Human Perception and Performance, 32*(1), 104–119. https://doi.org/10.1037/0096-1523.32.1.104.

Connine, C. M., Mullenix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1084–1096. https://doi.org/10.1037/0278-7393.16.6.1084

Connine, C. M., Ranbom, L. J., & Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Perception & Psychophysics, 70*(3), 403–411. https://doi.org/10.3758/PP.70.3.403

Connine, C. M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(1), 81–94. https://doi.org/10.1037/0278-7393.19.1.81

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Cruttenden, A. (2014). *Gimson’s pronunciation of English*. Routledge.

Dallas, B., McCarthy, A., & Long, G. (2016). Examining the educational benefits of and attitudes toward closed-captioning among undergraduate students. *Journal of the Scholarship of Teaching and Learning, 16*(2), 56–71. http://dx.doi.org/10.14434/josotl.v16i2.19267

Davis, R. (1998). *Running shoes*. http://www.esl-lab.com/running-shoes/runningshoesrd1.htm

Drijvers, L., Mulder, K., & Ernestus, M. (2016). Alpha and gamma band oscillations index differential processing of acoustically reduced and full forms. *Brain and Language, 153-154*, 27–37. https://doi.org/10.1016/j.bandl.2016.01.003

Duanmu, S. (2007). *The phonology of standard Chinese* (2nd ed.). Oxford University Press.

Education Bureau. (2004). *CDC english language curriculum guide (primary 1–6)*. The Government of the Hong Kong Special Administrative Region.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua, 142*, 27–41. https://doi.org/10.1016/j.lingua.2012.12.006

Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language, 81*(1), 162–173. https://doi.org/10.1006/brln.2001.2514

Field, A. P. (2013). *Discovering statistics using SPSS* (4th ed.). Sage.

Gernsbacher, M. A. (2015). *Video captions benefit everyone*. *Policy Insights from the Behavioral and Brain Sciences, 2*, 195–202. https://doi.org/10.1177/23773221502130

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Grgurović, M., & Hegelheimer, V. (2007). Help options and multimedia listening: Students’ use of subtitles and the transcript. *Language Learning & Technology, 11*(1), 45–66.

Guillory, H. G. (1998). The effects of keyword captions to authentic French video on learner comprehension. *CALICO Journal, 15*(1–3), 89–108.

Huang, H.-C., & Eskey, D. E. (2000). The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students. *Journal of Educational Technology Systems, 28*, 75–96.

Hulstijn, J. H. (2003). Connectionist models of language processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning, 16*(5), 413–425.

Johnson, K. (2004). Massive reduction in conversational American English. In K. Yoneyama & K. Maekawa (Eds.), *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium* (pp. 29–54). Tokyo, Japan: The National International Institute for Japanese Language.

Kuo, F.-L. (2012). Factors affecting Chinese EFL learners’ spoken word recognition. *NCUE Journal of Humanities, 11*(16), 98–106. https://doi.org/10.5539/ass.v11n16p98

Leon-Carrion, J., García-Orza, J., & Pérez-Santamaría, F. J. (2004). Development of the inhibitory component of the executive functions in children and adolescents. *International Journal of Neuroscience, 114*(10), 1291–1311. https://doi.org/10.1371/journal.pone.0077770

Liang, D. (2015). *Chinese learners’ pronunciation problems and listening difficulties in English connected speech*. *Asian Social Science, 11*(16), 98–106. https://doi.org/10.5539/ass.v11n16p98

Liao, C. Y.-W., & Yeldham, M. (2015). Taiwanese high school EFL teachers’ perceptions of their listening instruction. *The Asian Journal of Applied Linguistics, 2*(2), 92–101.
Lukas, S., Philipp, A. M., & Koch, I. (2010). Switching attention between modalities: Further evidence for visual dominance. *Psychological Research, 74,* 255–267. https://doi.org/10.1007/s00426-009-0246-y

Luo, L., & Lin, C.-T. (2012). The effects of captions in teenagers’ multimedia L2 learning. *Recall, 24*(2), 188–208. https://doi.org/10.1017/S0958344012000067

Maleki, A., & Rad, M. S. (2011). The effect of visual and textual accompaniments to verbal stimuli on the listening comprehension test performance of Iranian high and low proficient EFL learners. *Theory and Practice in Language Studies, 1*(1), 28–36.

Mao, H.-Z., & Chen, H.-Y. (2013). Exploring elision of schwa of /ə/ in English utterances by C & U English Majors. *International Journal of Applied Linguistics & English Literature, 2*(1), 117–125. https://doi.org/10.7575/ijael.v2n1p117

Markham, P. L., & Peter, L. (2003). The influence of English language and Spanish language captions on foreign language listening/reading comprehension. *Journal of Educational Technology Systems, 31*(3), 331–341. https://doi.org/10.2190/BHUH-420B-FE23-ALAO

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical access. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148–172). MIT Press.

Mattys, S. L., & Melhorn, J. F. (2007). Sentential, lexical, and acoustic effects on the perception of word boundaries. *Journal of the Acoustical Society of America, 122*(1), 554–567. https://doi.org/10.1121/1.2735105

Mitterer, H., & Russell, K. (2013). How phonological reductions sometimes help the listener. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(3), 977–984. https://doi.org/10.1037/a0029196

Pitt, M. A., & Samuel, A. G. (1995). Lexical and sublexical feedback in auditory word recognition. *Cognitive Psychology, 29*(2), 149–188. https://doi.org/10.1006/cogp.1995.1014

Pujola, J.-T. (2002). CALLing for help: Researching language learning strategies using help facilities in a web-based multimedia program. *Recall, 14*(2), 235–262. https://doi.org/10.1017/S0958344002000423

Ranbom, L. J., & Connine, C. M. (2007). Lexical representation of phonological variation in spoken word recognition. *Journal of Memory and Language, 57*(2), 273–298. https://doi.org/10.1016/j.jml.2007.04.001

Roach, P. (2008). *English phonetics and phonology: A practical course.* Foreign Language Teaching and Research Press.

Schnottz, W., & Kurschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review, 19,* 469–508. https://doi.org/10.1007/s10648-007-9053-4

Setter, J., Mok, P., Low, E. L., Zuo, D., & Tan, A. (2014). Word juncture characteristics in world Englishes: A research report. *World Englishes, 33,* 278–291.

Shockey, L. (2003). *Sound patterns of spoken English.* Blackwell.

Vandergrift, L. (2011). Second language listening: Presage, process, product, and pedagogy. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 455–471). Routledge.

Vanderplank, R. (2016). “Effects of” and “effects with” captions: How exactly does watching a TV programme with same-language subtitles make a difference in language learners? *Language Teaching, 49*(2), 235–250.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374–408. https://doi.org/10.1006/jmla.1998.2618

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology, 14*(1), 65–86. http://llt.msu.edu/vol14num1/winkekasssydorenko.pdf

Wong, S. W. L., Dealey, J., Mok, P., & Leung, V. W. -H. (in press). Production of English connected speech phonological processes: An assessment of Cantonese ESL learners’ difficulties in obtaining native-like speech. *The Language Learning Journal.* https://doi.org/10.1080/09571736.2019.1642372

Wong, S. W. L. Mok, P. P. K., Chung, K. K. -H., Leung, V. W. H., Bishop, D. V. M., & Chow, B. W. -Y. (2017a). Perception of native English reduced forms in Chinese learners: Its role in listening comprehension and its phonological correlates. *TESOL Quarterly, 51*(1), 7–31. https://doi.org/10.1002/tesq.273

Wong, S. W. L., Tsui, J. K. Y., Chow, B. W. -Y., Leung, V. W. H., Mok, P., & Chung, K. K. -H., (2017b). Perception of native English reduced forms in adverse environments by Chinese undergraduate students. *Journal of Psycholinguistic Research, 46*(5), 1149–1165. doi: 10.1007/s10936-017-9486-y

Yang, J. C., & Chang, P. (2014). Captions and reduced forms instruction: The impact on EFL students’ listening comprehension. *Recall, 26*(1), 44–61. https://doi.org/10.1017/S095834401300219

Yeldham, M. (2018). Viewing L2 captioned videos: What’s in it for the listener? *Computer Assisted Language Learning, 31*(4), 367–389. https://doi.org/10.1080/09588221.2017.1406956