Published in final edited form as:
Hear Res. 2008 August ; 242(1-2): 132–140. doi:10.1016/j.heares.2007.12.010.

Spectral and temporal cues for speech recognition: Implications for auditory prostheses

Li Xu[a,b,*] and **Bryan E. Pfingst**[b]

[a]*School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio 45701, USA*

[b]*Kresge Hearing Research Institute, Department of Otolaryngology, University of Michigan, Ann Arbor, Michigan 48109-5506, USA*

Abstract

Features of stimulation important for speech recognition in people with normal hearing and in people using implanted auditory prostheses include spectral information represented by place of stimulation along the tonotopic axis and temporal information represented in low-frequency envelopes of the signal. The relative contributions of these features to speech recognition and their interactions have been studied using vocoder-like simulations of cochlear implant speech processors presented to listeners with normal hearing. In these studies, spectral/place information was manipulated by varying the number of channels and the temporal-envelope information was manipulated by varying the lowpass cutoffs of the envelope extractors. Consonant and vowel recognition in quiet reached plateau at 8 and 12 channels and lowpass cutoff frequencies of 16 Hz and 4 Hz, respectively. Phoneme (especially vowel) recognition in noise required larger numbers of channels. Lexical tone recognition required larger numbers of channels and higher lowpass cutoff frequencies. There was a tradeoff between spectral/place and temporal-envelope requirements. Most current auditory prostheses seem to deliver adequate temporal-envelope information but the number of effective channels is suboptimal, particularly for speech recognition in noise, lexical tone recognition and music perception.

Keywords

speech recognition; spectral cues; temporal cues; cochlear implants

1. Introduction

Speech signals contain redundant cues that can be used for phoneme, word, and sentence recognition. Many studies have shown that dramatic reduction of the information in the speech signal does not cause significant reduction in recognition performance (e.g., Remez et al., 1981; Van Tasell et al., 1987; ter Keurs et al. 1992, 1993; Shannon et al., 1995). The reasonably good speech recognition outcomes of multichannel cochlear implant users provide an excellent example of how resilient speech recognition is under the conditions of reduced cues, at least in quiet. In current multichannel cochlear implants, speech signals are divided into a number

*Corresponding author. Address: School of Hearing, Speech and Language Sciences, Ohio University, Athens, Ohio 45701, USA. Email: xul@ohio.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

of frequency-specific components through a bank of bandpass filters. The envelope of each frequency band is extracted and then used to amplitude modulate an electric pulse train that is delivered to a more or less appropriate cochlear location along the electrode array. Therefore, representation of the spectral detail in the speech signal depends on the number of frequency bands in the processor. The temporal information is represented in the modulating envelopes. In this paper, we will treat the number of frequency channels as equivalent to spectral cues and the lowpass cutoff frequency of the temporal envelopes as equivalent to temporal cues. However, the details of the temporal envelopes also depend on the number of channels in the processor. As the number of channels increases, the temporal information in the envelope for each channel becomes restricted to narrower bands and the overall presentation of spectral-temporal information for the whole processor becomes more detailed.

The spectral resolution required for speech recognition has been studied using a spectral smearing technique (e.g., Villchur, 1977; ter Keurs et al., 1992, 1993; Baer and Moore, 1993, 1994; Boothroyd et al., 1996). It has been found in those studies that a reduced spectral resolution has detrimental effects on speech recognition. However, the effects of spectral smearing on speech recognition are minimal in quiet conditions, even for smearing that simulates auditory filters six times broader than normal (Baer and Moore, 1993, 1994). This result suggests that five or so bands of frequency information would be sufficient for speech recognition in quiet. Hill et al. (1968) used a vocoder technique invented by Homer Dudley (Dudley, 1939) and showed that with 6 to 8 channels of spectral information, the subjects obtained ~70% correct phoneme recognition. Using well-trained subjects, Shannon et al. (1995) showed that 4 channels are sufficient to achieve good (i.e., ≥ 85% correct) consonant, vowel, and sentence recognition in quiet. Today, there is converging evidence that 4 to 16 channels are necessary to achieve excellent speech recognition depending on the difficulties of the speech materials used (e.g., vowel, consonant, or sentence) and listening conditions (e.g., quiet or noise) (Shannon et al., 1995, 2004; Dorman et al., 1997; Loizou et al., 1999; Zeng, et al., 2005).

Other studies have focused on the effects of temporal cues on speech recognition. Rosen (1992) partitioned the temporal information in speech signals into three categories: (1) envelope (2–50 Hz), (2) periodicity (50–500 Hz), and (3) fine structure (500–10000 Hz). Most cochlear implant users cannot perceive differences in temporal information above about 300 Hz, so the high-frequency fine structure must be represented in a place code. Therefore, the temporal information of speech signal is hence referred to as the slow-varying envelope and periodicity cues. Van Tassel et al. (1987) studied consonant recognition using a one-channel noise band modulated by the speech envelope that was lowpassed at 20, 200, or 2000 Hz. They found that consonant recognition significantly improved when the envelope bandwidth increased from 20 to 200 Hz but did not improve further when the envelope bandwidth increased from 200 to 2000 Hz. More recent research using the temporal envelope smearing technique and the vocoder technique has shown that English speech recognition only benefits from envelope frequencies below 16 to 20 Hz (Drullman et al., 1994a, 1994b; Shannon et al., 1995; van der Horst et al., 1999; Fu and Shannon, 2000). Zeng et al. (1999) smeared the temporal envelope based on the temporal modulation transfer function (TMTF) measured from patients with auditory neuropathy, a cohort of auditory disorders characterized by dyssynchrony in auditory transmission. When the temporal envelope smeared speech signals were presented to the normal-hearing listeners, the latter exhibited various amount of degradation in speech understanding. In fact, the speech recognition performance of the normal-hearing listeners listening to the temporal envelope smeared speech could accurately predict the speech recognition performance of the patients with auditory neuropathy from whom the amount of smearing was derived.

Contemporary multichannel cochlear implants can apparently provide the many spectral and temporal cues needed for reasonably good speech recognition performance in quiet. Nonetheless, some limitations in terms of spectral and temporal cues received by cochlear implant users have to be considered. A few studies have demonstrated that most patients with cochlear implants are able to utilize only a maximum of seven to eight independent spectral channels (Fishman et al., 1997; Fu et al., 1998; Friesen et al., 2001). The temporal information received by the cochlear implant users is equivalent to that for normal-hearing listeners as measured by many psychophysical tests, such as gap detection, temporal integration, and temporal modulation transfer functions (Shannon, 1983; 1992). However, due to the highly synchronized firing in response to electrical stimulation (Wilson et al., 1997; Rubinstein et al., 1999), the representation and consequently the perception of temporal cues are likely to be different from those in normal-hearing listeners. Therefore, caution should be exerted when interpreting data from studies using acoustic models in normal-hearing subjects in relation to the performance in cochlear implant users. Nonetheless, the acoustic-simulation data can be viewed as the potential best scenario for the performance in cochlear implant users.

Although speech recognition is remarkably good in most patients with cochlear implants, with average sentence recognition ≥ 80% correct in quiet (see Zeng, 2004 for a review), reduced speech recognition in noise is one of the major problems faced by most implant users. Skinner et al. (2002) reported that one third of the 62 postlingually-deafened cochlear implant users tested at a fairly moderate +10 dB signal-to-noise ratio scored below 75% correct in sentence recognition. Another problem that cochlear-implant patients face is poor perception of tonal information. This results in a diminished enjoyment of music (e.g., Fujita and Ito, 1999; Gfeller et al., 2002, 2007; Kong et al., 2004; see also McDermott, 2004 for a review) and it is a major problem for people who speak tone languages (Zeng, 1995; Huang et al., 1995, 1996; Sun et al., 1998; Wei et al., 2000, 2004, 2007; Lee et al., 2002; Ciocca et al., 2002; Wong and Wong, 2004). The problem with music and lexical tone perception in cochlear implant users appears to stem from the same mechanism, that is, a limited number of spectral channels and a lack of encoding of fine structure information in the current implant systems (Smith et al., 2002; Xu and Pfingst, 2003). Thus, there is much room for improvement for both speech and music perception in cochlear implant users.

Despite the above mentioned studies on either spectral or temporal information for speech recognition, few studies have examined the relative contributions of both spectral and temporal cues to speech recognition at the same time. Data on how these two cues interact and affect speech recognition are needed to inform the design of speech-processing strategies that will provide better speech recognition for the cochlear implant users. We have undertaken a series of studies in an attempt to elucidate the relative importance of spectral and temporal cues and their interactions for phoneme recognition in quiet and in noise and for lexical-tone recognition (Xu et al., 2002, 2005; Xu and Zheng, 2007). The sections below provide a concise summary of the results of those studies and discuss the implications of the results for auditory prostheses in light of recent literature.

2. Spectral and temporal cues for phoneme recognition in quiet

To study the relative contributions of spectral and temporal cues to phoneme recognition, we varied both the amount of spectral and temporal information in vocoder-processed syllables presented to normal-hearing listeners. Detailed accounts of the signal processing techniques of vocoders can be found in Schroeder (1966) and Shannon et al. (1995). In our study, the spectral information was manipulated by varying the number of channels in the noise-excited vocoder from 1 to 16 whereas the temporal information was manipulated by varying the lowpass cutoff of the envelope extractor from 1 to 512 Hz. The total of 8 different numbers of channels (i.e., 1, 2, 3, 4, 6, 8, 12, and 16) and 10 lowpass cutoff frequencies (i.e., 1, 2, 4, 8, 16, 32, 64, 128, 256, and 512 Hz) created 80 vocoder conditions. Since the overall bandwidth was fixed (150–5500 Hz), the channel bandwidth covaried with the number of channels. Therefore, we could not differentiate the effects of channels bandwidth and number of channels here. Phoneme recognition tests consisted of a consonant test using the 20 initial consonants in a /Ca/ context (Shannon et al., 1999) and a vowel test using the 12 vowels in a /hVd/ context (Hillenbrand et al., 1995). Seven normal-hearing, native English-speaking adult listeners participated in the phoneme recognition tests.

To facilitate the comparison of the effects of spectral and temporal cues on phoneme recognition, we plotted the mean recognition scores in a contour plot format in which the areas filled with a particular color represent the percent correct scores at a given number of channels (abscissa) and lowpass cutoff frequency (ordinate) (Fig. 1). Although consonant and vowel recognition both depended on spectral and temporal cues, the pattern of dependence on spectral and temporal cues was different for consonants versus vowels. Vowel recognition predominantly depended on the number of channels as long as the lowpass cutoff frequencies were greater than about 4 Hz. Consonant recognition required a temporal envelope bandwidth as high as 16 to 32 Hz. It is interesting to note that there was a tradeoff of the spectral and temporal cues in phoneme recognition. That is, to achieve a certain level of phoneme recognition, an enhanced spectral cue could compensate for an impoverished temporal cue. On the other hand, if spectral information was very limited, for example, with only one spectral channel of stimulation, the listeners resorted to use of temporal information as high as 256 Hz of lowpass cutoff frequency. The tradeoff in consonant recognition occurred when the number of channels was ≤ 8 and the LPF was ≤ 16 Hz. For vowel recognition, some tradeoff between the number of channels and lowpass cutoff frequencies was observed for lowpass cutoff frequency ≤ 4 Hz and number of channels ≥ 4.

Recently, Nie et al. (2006) studied the spectral and temporal cues in patients with cochlear implants. The spectral cue was manipulated by varying the number of channels and the temporal cue was manipulated by varying the pulse rate. Their results confirmed the tradeoff effects between the spectral and temporal cues for speech recognition in cochlear implant users. In that study a smaller number of channels combined with a higher rate of stimulation achieved similar speech recognition scores to a larger number of channels combined with a lower rate of stimulation. The tradeoff was found to be particularly strong for consonant recognition but not for vowel recognition, consistent with our data using the acoustic simulations. However, it should be noted that even though the range of number of channels tested was equivalent to that studied in the acoustic simulations, the pulse rates used ranged from 1000 to 4000 Hz, which should be sufficient to deliver temporal envelope information greater than 500 Hz (Wilson et al., 1997). Therefore, the benefit of using higher rate stimulation is not clear but it is probably not due to a better representation of the temporal envelope cues, since a lowpass cutoff as low as 16 Hz was found to be sufficient for speech recognition. Nie et al. (2006) speculated that the improvement of speech recognition using high-rate stimulation resulted from the restoration of the normal stochastic properties in the neural responses (Rubinstein and Hong, 2003; Litvak et al., 2003a, b, c). However, other studies in patients with cochlear implants produced inconsistent results related to high-rate stimulation. Friesen et al. (2005) tested speech recognition in quiet in a group of cochlear implant users using pulse rates ranging from 200 to 5000 Hz and number of channels ranging from 4 to 16. They did not find a tradeoff between number of channels and stimulation rate. In fact, they found little difference in speech recognition using different stimulation rates and only 4-channel processors produced a significantly poorer performance than that found with a larger number of channels. Note that changing pulse rate is not equivalent to changing the low-pass cutoff frequency of the envelope extractor. The effects of pulse rate are complicated by variation in the neural responses to the rate.

3. Spectral and temporal cues for phoneme recognition in noise

Does speech recognition in noise require more spectral and temporal information? What are the relative contributions of the spectral and temporal cues for speech recognition in noise? Xu and Zheng (2007) attempted to address these questions using the technique of acoustic simulations of cochlear implants as described above. In that study, ten normal-hearing, native English-speaking adult listeners were recruited to participate in the phoneme recognition tests. All consonant (Shannon et al., 1999) and vowel tokens (Hillenbrand et al., 1995) were processed using the noise-excited vocoder as described above. Before the vocoder processing, the speech signal was mixed digitally with the speech-shaped noise (Nilsson et al., 1994) with signal-tonoise ratios (SNR) of +6 dB and 0 dB. The number of channels was varied between 2 and 32 (2, 4, 6, 8, 12, 16, 24, and 32). The lowpass cutoff frequency was varied between 1 and 512 Hz as in the experiments described above.

Fig. 2 shows the mean phoneme recognition scores of the ten normal-hearing listeners as a function of number of channels and lowpass cutoff frequency for the consonant (left column) and vowel (right column) tests under three conditions (quiet, SNR of +6 dB and SNR of 0 dB). The contour plot format is the same as Fig. 1 except for the abscissa scales.

To quantify the performance plateau, an exponential function, $y = a \cdot e^{b(x+c)} + d$, where y is the percent-correct scores and x is the number of channels or lowpass cutoff frequency, was used to fit each of the original group-mean score curves. The values of the parameters of the exponential function were derived based on the method of ordinary least squares. From each of the fitted curves, we determined a knee point, defined as the number of channels, or lowpass cutoff frequency at which 90% of the performance plateau was reached. The black lines in Fig. 2 show the knee-point data with the vertical lines representing the number of channels at which the recognition performance became asymptotic and the horizontal line representing the lowpass cutoff frequencies at which the recognition performance became asymptotic. For both quiet and noise conditions, the number of channels required for consonant recognition to reach plateau was around 12, indicating that an increase of spectral information beyond 12 channels could not improve consonant recognition in noise. On the other hand, the number of channels required for vowel recognition to reach plateau was around 12 in quiet and between 16 and 24 in the noise conditions, indicating that vowel recognition in noise benefited from the increased spectral information. In terms of temporal information, the lowpass cutoff frequencies for the performance plateau were between 8 and 16 Hz for consonant recognition and around 4 Hz for vowel recognition in both quiet and noise conditions.

The two lines also divide each of the contour plots in Fig. 2 into four quadrants. Of particular interest is the lower-left quadrant in which there was a tradeoff between the spectral and the temporal cues for phoneme recognition. It was evident that the tradeoff existed for consonant and vowel recognition in both quiet and noise conditions. The tradeoff for consonant recognition occurred with the number of channels ≤ 12 and the lowpass cutoff frequency ≤ 16 Hz, whereas the ranges of tradeoff between the spectral and temporal cues for vowel recognition were quite limited.

Several studies have found that a larger number of channels was required in noise conditions than in quiet for cochlear implant users to achieve the plateau performance (Fu et al., 1998; Friesen et al., 2001). Our results suggest that the improvement of speech recognition in noise with a larger number of channels is probably due to the improvement in vowel recognition as opposed to consonant recognition. The utility of temporal information for speech recognition in noise has not been extensively explored. Our results demonstrated that temporal cues contributed to phoneme recognition similarly in both quiet and noise conditions. As illustrated in Fig. 2, the lowpass cutoff frequencies required for plateau performance were almost constant in the noise conditions for consonant (16 Hz) and vowel (4 Hz) recognition as well as in quiet condition (for speech recognition in quiet, see also Fu and Shannon, 2000 and Xu et al., 2005). We speculate that there are several possible explanations for the limitations of the auditory system to use more temporal information for speech recognition in noise. First, much of the temporal envelope information important for phoneme recognition resides in the low frequency region (see also van der Horst et al., 1999). Therefore, additional temporal envelope cues do not provide further improvement in phoneme recognition. Second, noise might obscure temporal envelope cues, making the cues less effective. We used a steady-state speech-shape noise in the study reported above. Other types of noise, such as temporal-modulated noise or single-talker noise (Qin and Oxenham, 2003; Stickney et al., 2004), might produce even greater interferences of the temporal envelope, leading to a less effective use of the temporal envelope cues.

4. Spectral and temporal cues for lexical-tone recognition

Pitch perception is particularly poor in current cochlear implant users. This poses a special challenge for cochlear implant users who speak tone languages. Fig. 3 shows the waveforms and spectrograms of a Mandarin Chinese syllable /shi/ that was spoken by a female adult in the four Mandarin tone patterns. The tone patterns, as defined by the fundamental frequency (F0) contours, vary in the following four ways: (1) flat, (2) rising, (3) falling and then rising, and (4) falling. The four tones of /shi/ in Mandarin Chinese can mean (1) lion [狮], (2) eat [食], (3) history [史], and (4) yes [是], respectively. Since F0 and its harmonics in speech signals, which are the primary acoustic cues for tone perception under normal-hearing conditions, are not explicitly coded in the speech processing strategies of cochlear implants, we undertook a study to examine whether the temporal and spectral cues that are present in cochlear implant stimulation are usable for lexical-tone recognition (Xu et al., 2002).

The same experimental paradigm as described above was used for the study of lexical-tone recognition. Four normal-hearing, native Mandarin-speaking adult listeners participated in this part of the experiments. Speech materials consisted of 10 monosyllables (i.e., /ma/, /ji/, /wan/, /yi/, /fu/, /xian/, /qi/, /yan/, /yang/, /xi/) spoken in four Mandarin tones by a male and a female adult speaker, resulting in a total of 80 tone tokens (Xu et al., 2002). During recording, care was taken to ensure that the four tone tokens of the same monosyllable were of equal duration. The durations of naturally spoken, isolated Mandarin Chinese words vary systematically, with tone 3 typically being the longest and tone 4 the shortest. These duration cues have been shown to assist in tone recognition (Whalen and Xu, 1992; Fu and Zeng, 2000; Xu et al., 2002). However, syllable duration is not a reliable cue in connected speech in everyday situations. Duanmu (2002) has summarized a few studies on duration of tones and has found that the differences in duration for Mandarin tones are rather small, all within 10% among different tones. Therefore, we used tone tokens with equal duration so that we could focus on the study of tone recognition with the potentially confounding variable of duration brought under control.

All tone tokens were processed using the noise-excited vocoder described above. The number of channels was varied from 1 to 12 (i.e., 1, 2, 3, 4, 6, 8, 10, and 12) whereas the lowpass cutoff frequency was covaried between 1 and 512 Hz (i.e., 1, 2, 4, 8, 16, 32, 64, 128, 256, and 512 Hz). The subjects performed the four-alternative forced-choice tone recognition tests of the 80 combinations of number of channels and lowpass cutoff frequency in a random order. Fig. 4A summarized the mean tone recognition scores across the four subjects. Tone recognition depended on both number of channels and lowpass cutoff frequency. The performance increased as the number of channels increased from 1 to 12 and as lowpass cutoff frequency increased from 1 to 256 Hz. Those ranges of number of channels and lowpass cutoff frequency were where the tradeoff of spectral and temporal cues occurred for tone recognition. Compared to English phoneme recognition (Fig. 1), the overall tone recognition performance was remarkably lower. The best performance with 12 channels (the largest number tested) and a lowpass cutoff frequency ≥ 256 Hz was around 75% correct. Note that the chance performance for tone recognition tests was 25% correct, as opposed to 5% and 8.3% correct for consonant and vowel tests. Another dramatic difference between tone recognition and phoneme recognition was the stronger dependence on temporal information for tone recognition. At the higher lowpass cutoff frequencies (e.g., ≥64 Hz), periodicity cues that are directly related to the voice pitch may have become available to assist tone recognition by the listeners.

Fig. 4B, C, and D show the waveforms and spectrograms of the vocoder processed syllables in Mandarin tones with numbers of channels of 12, 2, and 4 and the lowpass cutoff frequencies of 512, 2, and 16 Hz, respectively. These graphs help to provide some descriptive accounts of the results. Temporal envelopes have been shown to convey Mandarin tone information (Whalen and Xu, 1992). In our examples, they are better preserved at lowpass cutoff frequencies > 2 Hz (Fig. 4B and D). The spectrograms show that the F0 and its harmonics are absent in all tokens. However, with as few as four channels, the formants (i.e., F1 and F2) can roughly be discerned (Fig. 4D). With 12 channels, the formants become fairly clear (Fig. 4B). Higher lowpass cutoff frequencies provide richer and richer temporal information across different channels (Fig. 4B and D). This makes the temporal-spectral patterns of the four tone tokens somewhat distinguishable.

In another experiment, we used vocoder-processed frequency-modulated pulse trains as stimuli in tone recognition tests and found that the subjects required 30 to 40 spectral channels to achieve recognition performances near 75% correct (see Fig. 9 in Xu et al., 2002). When the number of channels is > 30, some of the harmonics in the speech signals can be resolved and thus voice pitch information may become available to listeners. In a recent study of spectral and temporal cues for Mandarin Chinese tones, Kong and Zeng (2006) found that 32 spectral bands were needed to achieve performance similar to that obtained with the original stimuli. Another interesting finding in Kong and Zeng (2006) is that there was an interaction among the spectral cues, temporal cues and noise levels in tone recognition. In quiet conditions, tone recognition performance with 1-band 500-Hz lowpass cutoff stimulation was better than that with the 8-band 50-Hz lowpass cutoff stimulation. In the noise condition, a reversed pattern was observed. This suggests that enhanced spectral cues are important for tone recognition in noise, consistent with our findings with vowel recognition in noise (Xu and Zheng, 2007).

A few studies have tested the importance of spectral and/or temporal cues for tone recognition in native tone language speaking cochlear implant users. Wei et al. (2004) tested tone recognition in five Nucleus-22 users and showed a clear dependence of tone recognition on number of channels. The performance improved from about 30% correct with one channel to about 70% correct with ten channels and reached plateau at ten channels. Liu et al. (2004) found that tone recognition in six Mandarin-speaking children who use the ACE (advanced combined encoder) strategy in their CI24 devices was not seriously affected when the following electrodes were eliminated: (1) half of the nonadjacent 22 electrodes, (2) six basal electrodes, and (3) 6 apical electrodes. However, with a map using only the six apical electrodes, tone recognition decreased by 30 percentage points from the original map. Fu et al. (2004) tested tone recognition in nine Mandarin-speaking CI24 users with various stimulation rates. All subjects were ACE users and did not demonstrate any significant differences in tone recognition when the stimulation rates were varied from 900 to 1800 Hz. When the speech processing strategy was converted to a slower rate SPEAK (spectral peak) strategy with a 250-Hz stimulation rate, all subjects showed a significant decrease in tone recognition. It was not clear whether the results were due to the fact that the subjects were all ACE users and the SPEAK strategy was new to them or whether the ACE strategy actually provided more temporal information to the subjects. Interestingly, when fitted with the CIS (continuous interleaved sampling) using 12, 8, 6, and 4 electrodes coupled with stimulation rates of 1200, 1800, 2400, and 3600 Hz, respectively, the subjects all showed equivalent tone recognition performance across different maps. It appeared that there was a tradeoff between the number of electrodes and the stimulation rate for tone recognition as was found by Nie et al. (2006) for vowel recognition.

5. General Discussion

To optimize current cochlear implant systems for the users, it is important to understand the mechanisms by which subjects perceive speech signals under conditions where the signals are degraded. It is also important to learn which aspects of the speech cues have the potential to improve speech recognition if enhanced and whether strengthening one type of cue can compensate for weaknesses in other types. In reviewing a series of studies using the acoustic simulations of cochlear implants, we have examined the relative contributions of spectral and temporal cues to phoneme recognition in quiet and in noise and to lexical-tone recognition. We varied the number of channels to control the amount of spectral information in the speech signal and at the same time varied the lowpass cutoff frequencies of the envelope extractors to control the range of temporal-envelope information in the signal. Normal-hearing adult subjects participated in the perceptual tests. It was found that both temporal and spectral cues were important for English phoneme recognition in quiet and in noise. To reach performance plateau for consonant and vowel recognition in quiet, the numbers of channels required were 8 and 12, respectively and the lowpass cutoff frequencies required were 16 and 4 Hz, respectively. Interestingly, consonant recognition in noise did not benefit from additional spectral or temporal cues beyond those required for plateau performance in quiet. Vowel recognition in noise, on the other hand, reached performance plateau with a larger number of channels (i.e., an increased spectral resolution) but not a higher lowpass cutoff frequency than was required in quiet. In both noise and quiet conditions, there was a tradeoff between the temporal and spectral cues for English phoneme recognition. Limited data from the cochlear implant users appear to be consistent with our finding in that more spectral channels are necessary for speech recognition in noise (Fishman et al., 1997; Fu et al., 1998; Friesen et al., 2001). Therefore, to improve speech recognition, especially in noise, efforts should be concentrated on providing a larger number of effective channels in the cochlear implant systems.

Performance for Mandarin-Chinese tone recognition was remarkably poorer than that for English phoneme recognition under comparable conditions. The lowpass cutoff frequency required for asymptotic performance of tone recognition was as high as 256 Hz, much higher than that required for English phoneme recognition (16 Hz for consonant and 4 Hz for vowel recognition). Xu et al., (2002) found that tone recognition performance improved as a function of number of channels and had not reached plateau at 12 channels. Another study indicated that as many as 32 channels were necessary for lexical-tone recognition to be close to normal listening conditions (Kong and Zeng, 2006). Xu et al. (2002) also found that, as in English phoneme recognition, there was a tradeoff between the temporal and spectral cues for lexical tone recognition. Efforts aimed at enhancing the temporal envelope cues for tone recognition or pitch perception using acoustic simulations or actual cochlear implants have yielded only modest success (e.g., Geurts and Wouters, 2001; Green et al., 2004, 2005; Hamilton et al., 2007; Laneau et al., 2006; Luo and Fu, 2004; Vandali et al., 2005, 2007). On the other hand, few studies have attempted to manipulate the spectral cues to improve tone recognition or pitch perception (e.g., Geurts and Wouters, 2004). With the advent of current-steering stimulation in clinical speech processors, a large number of channels can potentially be achieved. We are just beginning to test whether such a stimulation strategy will produce better lexical-tone perception in the implant users (Xu et al., 2007). It is important to make a distinction between the number of channels stimulated and the number of effective channels. Current technology can dramatically increase the number of stimulation sites by using either densely packed electrodes or virtual channels. However, other limiting factors have to be taken into account before larger numbers of effective channels can be achieved. Those factors include channel interaction and neural survival patterns.

There is still much to learn about spectral and temporal cues for speech recognition. For example, much more data are needed to understand the confusion patterns of phoneme recognition in noise under conditions of reduced spectral and temporal cues. Additionally, our understanding of how spectral and temporal cues are used by patients with sensorineural hearing loss is limited (e.g., Turner et al., 1999; Baskent, 2006). Future research is warranted to elucidate the interactions of the two cues in patients with cochlear loss. Finally, the role of cognitive function, particularly working memory, in speech recognition studies using the vocoder technique remains to be determined.

6. Concluding Remarks

Our results for normal-hearing subjects listening to vocoded speech demonstrated that both temporal and spectral cues were important for phoneme recognition in quiet and in noise. Plateau performance for consonant and vowel recognition in quiet was reached when the numbers of channels were 8 and 12, respectively and the lowpass cutoff frequencies were 16 Hz and 4 Hz, respectively. In noise conditions, vowel recognition benefited from increased spectral resolution, reaching asymptote at 16–24 channels. For lexical-tone recognition in quiet, the lowpass cutoff frequency required for asymptotic performance was as high as 256 Hz. Tone recognition performance in quiet did not reach plateau by 12 channels, the highest number tested. There was a tradeoff between the temporal and spectral cues for phoneme and lexical-tone recognition. Since current implant systems appear to deliver adequate temporal envelope information, to improve speech recognition in noise and lexical-tone recognition, efforts should be concentrated on providing a larger number of effective channels in the cochlear implant systems and other auditory prostheses.

ACKNOWLEDGEMENTS

We acknowledge the technical assistance from Cathy Thompson, Yuhjung Tsai, and Yunfang Zheng. Ning Zhou provided valuable comments on an earlier version of the manuscript. Supported by NIH Grants R01 DC03808, T32 DC00011, F32 DC00470, R03 DC006161, P30 DC05188 and Ohio University research awards.

References

Baer T, Moore BCJ. Effects of spectral smearing on the intelligibility of sentences in noise. J. Acoust. Soc. Am 1993;94:1229–1241.

Baer T, Moore BCJ. Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. J. Acoust. Soc. Am 1994;95:2277–2280. [PubMed: 8201124]

Baskent D. Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels. J. Acoust. Soc. Am 2006;120:2908–2925. [PubMed: 17139748]

Boersma, P.; Weenink, D. Praat: Doing phonetics by computer (Version 4.6.09). 2007. Retrieved July 10, 2007, from http://www.praat.org/

Boothroyd A, Mulhearn B, Gong J, Ostroff J. Effects of spectral smearing on phoneme and word recognition. J. Acoust. Soc. Am 1996;100:1807–1818. [PubMed: 8817914]

Ciocca V, Francis AL, Aisha R, Wong L. The perception of Cantonese lexical tones by early-deafened cochlear implantees. J. Acoust. Soc. Am 2002;111:2250–2256. [PubMed: 12051445]

Dorman MF, Loizou PC, Rainey D. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. J. Acoust. Soc. Am 1997;102:2403–2411. [PubMed: 9348698]

Drullman R, Festen JM, Plomp R. Effect of temporal envelope smearing on speech perception. J. Acoust. Soc. Am 1994a;95:1053–1064. [PubMed: 8132899]

Drullman R, Festen JM, Plomp R. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am 1994b;95:2670–2680. [PubMed: 8207140]

Duanmu, S. The phonology of standard Chinese. Oxford: Oxford University Press; 2002.

Dudley H. The vocoder. Bell Labs Rec 1939;17:122–126.

Fishman KE, Shannon RV, Slattery WH. Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. J. Speech Lang. Hear. Res 1997;40:1201–1215. [PubMed: 9328890]

Friesen LM, Shannon RV, Baskent D, Wang X. Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am 2001;110:1150–1163. [PubMed: 11519582]

Friesen LM, Shannon RV, Cruz RJ. Effects of stimulation rate on speech recognition with cochlear implants. Audiol. NeuroOtol 2005;10:169–184. [PubMed: 15724088]

Fujita S, Ito J. Ability of Nucleus cochlear implantees to recognize music. Ann. Otol. Rhinol. Laryngol 1999;108:634–640. [PubMed: 10435919]

Fu Q-J, Hsu C-J, Horng M-J. Effects of speech processing strategy on Chinese tone recognition by Nucleus-24 cochlear implant patients. Ear Hear 2004;25:501–508. [PubMed: 15599196]

Fu Q-J, Shannon RV. Effect of stimulation rate on phoneme recognition by Nucleus-22 cochlear implant listeners. J. Acoust. Soc. Am 2000;107:589–597. [PubMed: 10641667]

Fu Q-J, Shannon RV, Wang X. Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. J. Acoust. Soc. Am 1998;104:3586–3596. [PubMed: 9857517]

Fu Q-J, Zeng F-G. Identification of temporal envelope cues in Chinese tone recognition. Asia Pacific Journal of Speech, Language and Hearing 2000;5:45–57.

Gfeller K, Turner C, Woodworth G, Mehr M, Fearn R, Witt S, Stordahl J. Recognition of familiar melodies by adult cochlear implant recipients and normal hearing adults. Cochlear Implants International 2002;3:29–53. [PubMed: 18792110]

Gfeller K, Turner C, Oleson J, Zhang X, Gantz B, Froman R, Olszewski C. Accuracy of cochlear implant recipients on pitch perception, melody recognition, and speech reception in noise. Ear Hear 2007;28 (3):412–423. [PubMed: 17485990]

Geurts L, Wouters J. Coding of the fundamental frequency in continuous interleaved sampling processors for cochlear implants. J. Acoust. Soc. Am 2001;109:713–726. [PubMed: 11248975]

Geurts L, Wouters J. Better place-coding of the fundamental frequency in cochlear implants. J. Acoust. Soc. Am 2004;115:844–852. [PubMed: 15000196]

Green T, Faulkner A, Rosen S. Enhancing temporal cues to voice pitch in continuous interleaved sampling cochlear implants. J. Acoust. Soc. Am 2004;116:2298–2310. [PubMed: 15532661]

Green T, Faulkner A, Rosen S, Macherey O. Enhancement of temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification. J. Acoust. Soc. Am 2005;118:375–385. [PubMed: 16119358]

Hamilton N, Green T, Faulkner A. Use of a single channel dedicated to conveying enhanced temporal periodicity cues in cochlear implants: Effects on prosodic perception and vowel identification. Int. J. Audiol 2007;46:244–253. [PubMed: 17487672]

Hill FJ, McRae LP, McClellan RP. Speech recognition as a function of channel capacity in a discrete set of channels. J. Acoust. Soc. Am 1968;44:13–18. [PubMed: 5659828]

Hillenbrand J, Getty LA, Clark MJ, Wheeler K. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am 1995;97:3099–3111. [PubMed: 7759650]

Huang TS, Wang NM, Liu SY. Tone perception of Mandarin-speaking postlingually deaf implantees using the nucleus 22-channel cochlear mini system. Ann. Otol. Rhinol. Laryngol. Suppl 1995;166:294–298. [PubMed: 7668677]

Huang TS, Wang NM, Liu SY. Nucleus 22-channel cochlear mini-system implantations in Mandarin-speaking patients. Am. J. Otol 1996;17:46–52. [PubMed: 8694134]

Kong Y-Y, Cruz R, Jones JA, Zeng F-G. Music perception with temporal cues in acoustic and electric hearing. Ear Hear 2004;25:173–185. [PubMed: 15064662]

Kong Y-Y, Zeng F-G. Temporal and spectral cues in Mandarin tone recognition. J. Acoust. Soc. Am 2006;120:2830–2840. [PubMed: 17139741]

Laneau J, Wouters J, Moonen M. Improving music perception with explicit pitch coding in cochlear implants. Audiol. NeuroOtol 2006;11:38–52. [PubMed: 16219993]

Lee KYS, Van Hasselt CA, Chiu SN, Cheung DMC. Cantonese tone perception ability of cochlear implant children in comparison with normal-hearing children. Int. J. Pediatr. Otorhinolaryngol 2002;63:137–147. [PubMed: 11955605]

Litvak L, Delgutte B, Eddington D. Improved neural representation of vowels in electric stimulation using desynchronizing pulse trains. J. Acoust. Soc. Am 2003a;114:2099–2111. [PubMed: 14587608]

Litvak LM, Delgutte B, Eddington DK. Improved temporal coding of sinusoids in electric stimulation of the auditory nerve using desynchronizing pulse trains. J. Acoust. Soc. Am 2003b;114:2079–2098. [PubMed: 14587607]

Litvak LM, Smith ZM, Delgutte B, Eddington DK. Desynchronization of electrically evoked auditory-nerve activity by high-frequency pulse trains of long duration. J. Acoust. Soc. Am 2003c;114:2066–2078. [PubMed: 14587606]

Liu T-C, Chen H-P, Lin H-C. Effects of limiting the number of active electrodes on Mandarin tone perception in young children using cochlear implants. Acta Otolaryngol 2004;124:1149–1154. [PubMed: 15768808]

Loizou PC, Dorman M, Tu Z. On the number of channels needed to understand speech. J. Acoust. Soc. Am 1999;106:2097–2103. [PubMed: 10530032]

Luo X, Fu Q-J. Enhancing Chinese tone recognition by manipulating amplitude envelope: Implications for cochlear implants. J. Acoust. Soc. Am 2004;116:3659–3667. [PubMed: 15658716]

McDermott H. Music perception with cochlear implants: A review. Trends in Amplification 2004;8:49–81. [PubMed: 15497033]

Nie K, Barco A, Zeng F-G. Spectral and temporal cues in cochlear implant speech perception. Ear Hear 2006;27:208–217. [PubMed: 16518146]

Nilsson M, Sali SD, Sullivan JA. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am 1994;95:1085–1099. [PubMed: 8132902]

Qin MK, Oxenham AJ. Effects of simulated cochlear implant processing on speech reception in fluctuating maskers. J. Acoust. Soc. Am 2003;114:446–454. [PubMed: 12880055]

Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science 1981;212:947–950. [PubMed: 7233191]

Rosen S. Temporal information in speech: Acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. London, Ser. B 1992;336:367–373. [PubMed: 1354376]

Rubinstein JT, Hong RS. Signal coding in cochlear implants: exploiting effects of electrical stimulation. Ann. Otol. Rhinol. Laryngol 2003;112:14–19. [PubMed: 12537052]

Rubinstein JT, Wilson BS, Finley CC, Abbas PJ. Pseudospontaneous activity: Stochastic independence of auditory nerve fibers with electrical stimulation. Hear Res 1999;127:108–118. [PubMed: 9925022]

Schroeder MR. Vocoders: Analysis and synthesis of speech. Proc. IEEE 1966;54:352–366.

Shannon RV. Multichannel electrical stimulation of the auditory nerve in man. I. basic psychophysics. Hear. Res 1983;11:157–189. [PubMed: 6619003]

Shannon RV. Temporal modulation transfer functions in patients with cochlear implants. J. Acoust. Soc. Am 1992;91:2156–2164. [PubMed: 1597606]

Shannon RV, Fu Q-J, Galvin J. The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. Acta Otolaryngol 2004:50–54.

Shannon RV, Jensvold A, Padilla M, Robert ME, Wang X. Consonant recordings for speech testing. J. Acoust. Soc. Am 1999;106:L71–L74. [PubMed: 10615713]

Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science 1995;270:303–304. [PubMed: 7569981]

Skinner MW, Arndt PL, Staller SJ. Nucleus 24 advanced encoder conversion study: Performance vs. preference. Ear Hear 2002;23:2S–25S. [PubMed: 11883765]

Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. Nature 2002;416:87–90. [PubMed: 11882898]

Stickney GS, Zeng F-G, Litovsky R, Assmann P. Cochlear implant speech recognition with speech maskers. Acoust. Soc. Am 2004;116:1081–1091.

Sun JC, Skinner MW, Liu SY, Wang FNM, Huang TS, Lin T. Optimization of speech processor fitting strategies for Chinese- speaking cochlear implantees. Laryngoscope 1998;108:560–568. [PubMed: 9546270]

ter Keurs M, Festen JM, Plomp R. Effect of spectral envelope smearing on speech reception I. J. Acoust. Soc. Am 1992;91:2872–2880. [PubMed: 1629480]

ter Keurs M, Festen JM, Plomp R. Effect of spectral envelope smearing on speech reception. II. J. Acoust. Soc. Am 1993;93:1547–1552. [PubMed: 8473608]

Turner CW, Chi S-L, Flock S. Limiting spectral resolution in speech for listeners with sensorineural hearing loss. J. Speech Lang. Hear. Res 1999;42:773–784. [PubMed: 10450899]

van der Horst R, Leeuw AR, Dreschler WA. Importance of temporal-envelope cues in consonant recognition. J Acoust Soc Am 1999;105:1801–1809. [PubMed: 10089603]

Van Tassel DJ, Soli SD, Kirby VM, Widin GP. Speech waveform envelope cues for consonant recognition. J. Acoust. Soc. Am 1987;82:1152–1161. [PubMed: 3680774]

Vandali AE, Sucher C, Tsang DJ, McKay CM, Chew JW, McDermott HJ. Pitch ranking ability of cochlear implant recipients: A comparison of sound-processing strategies. J Acoust. Soc. Am 2005;117:3126–3138. [PubMed: 15957780]

Vandali AE, Ciocca V, Wong LLN, Luk B, Ip VWK, Murray B, Yu HC, Chung I, Ng E, Yuen K. Pitch and tonal language perception in cochlear implant users. Conference on Implantable Auditory Prostheses Abst 2007:204.

Villchur E. Electronic models to simulate the effect of sensory distortions on speech perception by the deaf. J. Acoust. Soc. Am 1977;62:665–674. [PubMed: 903508]

Wei C-G, Cao K, Zeng F-G. Mandarin tone recognition in cochlear-implant subjects. Hear. Res 2004;197:87–95. [PubMed: 15504607]

Wei C, Cao K, Jin X, Chen X, Zeng F-G. Psychophysical performance and Mandarin tone recognition in noise by cochlear implant users. Ear Hear 2007;28:62S–65S. [PubMed: 17496650]

Wei WI, Wong R, Hui Y, Au DKK, Wong BYK, Ho WK, Tsang A, Kung P, Chung E. Chinese tonal language rehabilitation following cochlear implantation in children. Acta Otolaryngol 2000;120:218–221. [PubMed: 11603776]

Whalen DH, Xu Y. Information for Mandarin tones in the amplitude contour and in brief segments. Phonetica 1992;49:25–47. [PubMed: 1603839]

Wilson BS, Finley CC, Lawson DT, Zerbi M. Temporal representations with cochlear implants. Am. J. Otol 1997;18(6 Suppl):S30–S34. [PubMed: 9391587]

Wong AO, Wong LL. Tone perception of Cantonese-speaking prelingually hearing-impaired children with cochlear implants. Otolaryngol. Head Neck Surg 2004;130:751–758. [PubMed: 15195063]

Xu, L.; Han, D.; Liu, B.; Chen, X.; Kong, Y.; Liu, H.; Zheng, Y.; Zhou, N. Asia-Pacific Symposium on Cochlear Implants. Sydney, Australia: 2007. Lexical tone perception with HiResolution® 120 speech-processing strategy in Mandarin-speaking children.

Xu L, Pfingst BE. Relative importance of temporal envelope and fine structure in lexical-tone perception. J. Acoust. Soc. Am 2003;114:3024–3027. [PubMed: 14714781]

Xu L, Thompson CS, Pfingst BE. Relative contributions of spectral and temporal cues for phoneme recognition. J. Acoust. Soc. Am 2005;117:3255–3267. [PubMed: 15957791]

Xu L, Tsai Y, Pfingst BE. Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses. J. Acoust. Soc. Am 2002;112:247–258. [PubMed: 12141350]

Xu L, Zheng Y. Spectral and temporal cues for phoneme recognition in noise. J. Acoust. Soc. Am 2007;122:1758–1764. [PubMed: 17927435]

Zeng F-G. Cochlear implants in China. Audiology 1995;34:61–75. [PubMed: 8561684]

Zeng F-G. Trends in cochlear implants. Trends in Amplification 2004;8:1–34. [PubMed: 15247993]

Zeng F-G, Nie K, Stickney GS, Kong Y-Y, Vongphoe M, Bhargave A, Wei C, Cao K. Speech recognition with amplitude and frequency modulations. Proc. Natl. Acad. Sci. U.S.A 2005;102:2293–2298. [PubMed: 15677723]

Zeng F-G, Oba S, Garde S, Sininger Y, Starr A. Temporal and speech processing deficits in auditory neuropathy. NeuroReport 1999;10(16):3429–3435. [PubMed: 10599857]

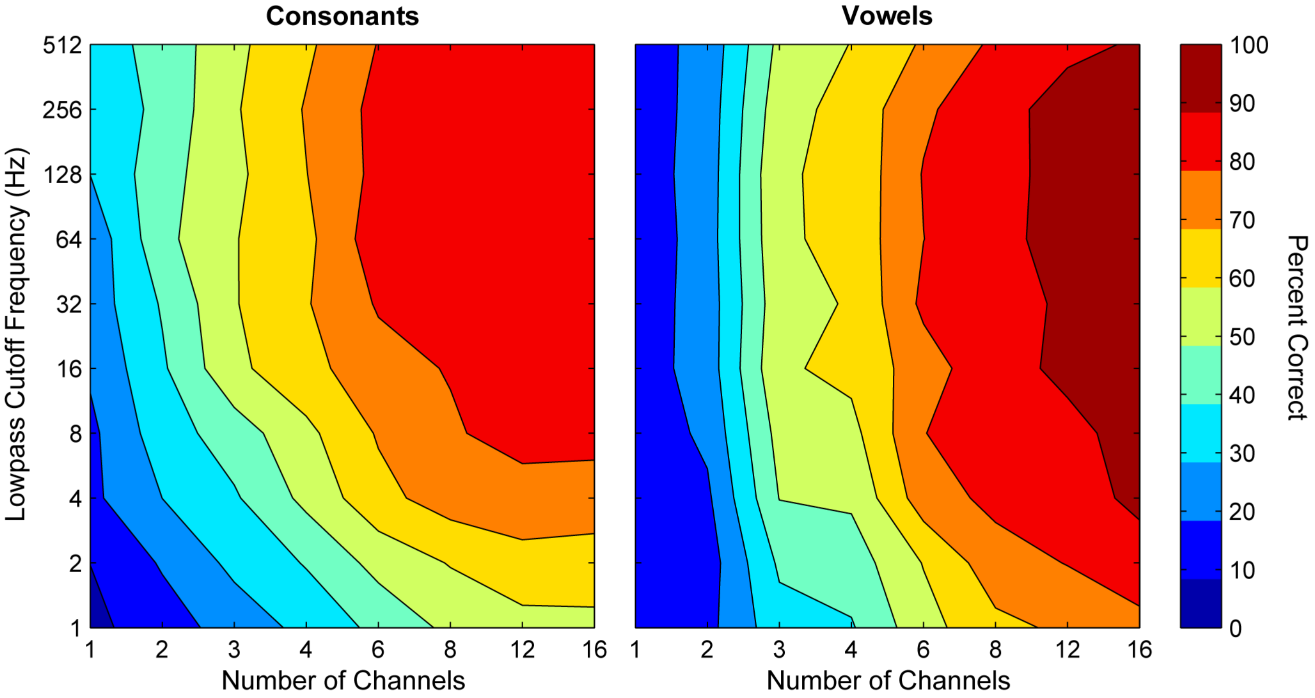

Fig. 1.

Mean phoneme recognition scores (percent correct) as a function of the number of channels and lowpass cutoff frequency. The left and right panels represent data for consonant and vowel recognition, respectively. In each contour plot, the area that is filled with a particular color represents the phoneme recognition score for a given number of channels (abscissa) and lowpass cutoff frequency (ordinate). The percent correct represented by the color is indicated by the bar on the right. Adapted from Xu et al. (2002) with permission from the Acoustical Society of America.

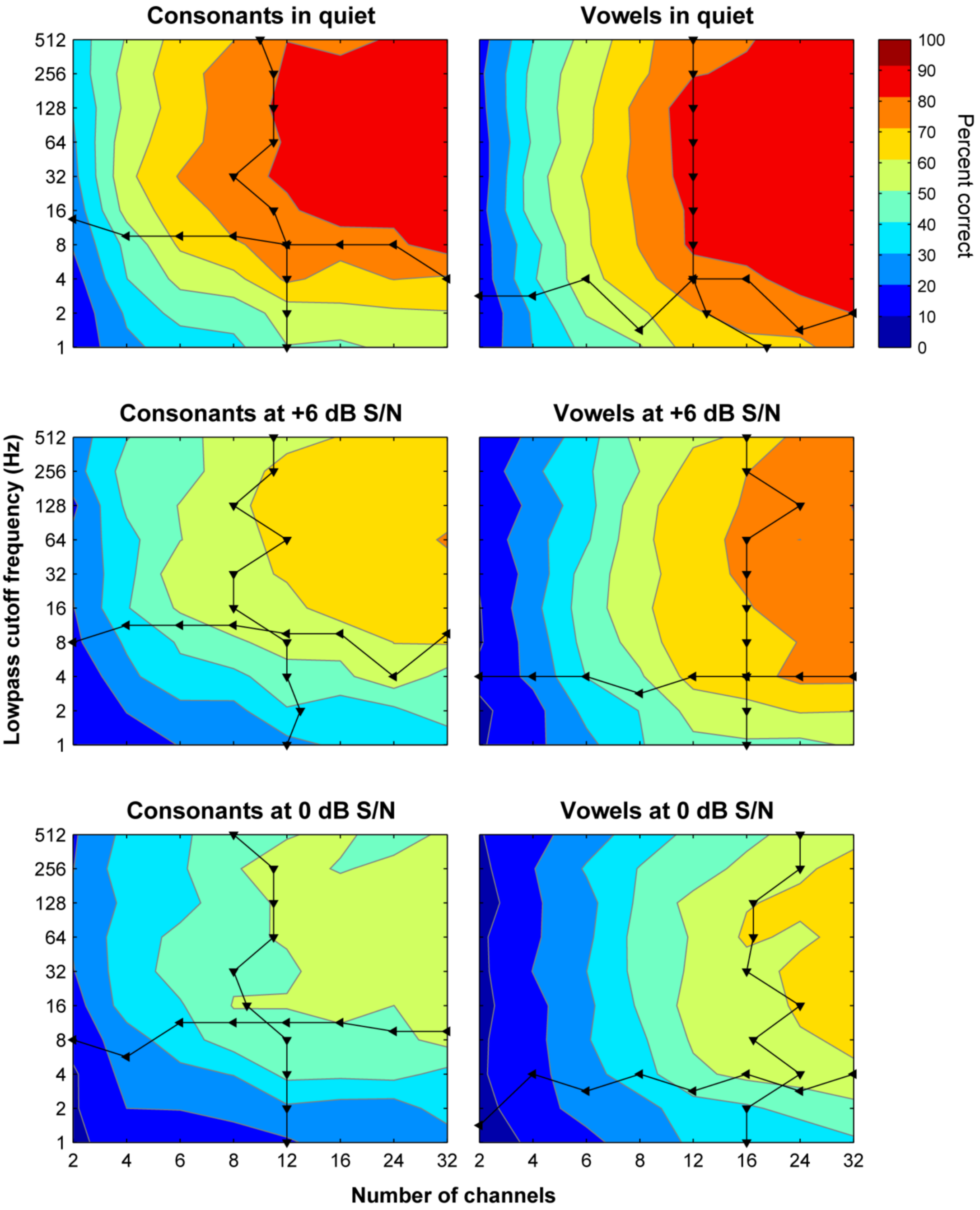

Fig. 2.

Group-mean phoneme recognition as a function of both number of channels (abscissa) and lowpass cutoff frequency (ordinate) under three conditions (top row: quiet; middle row: SNR of +6 dB; bottom row: SNR of 0 dB) for consonant (left) and vowel (right) tests. The vertical line and the symbol (▼) represent the knee points (i.e., the number of channels at which the recognition performance reached 90% of the performance plateau) using the corresponding lowpass cutoff frequency indicated on the ordinate. The horizontal line and the symbol (◀) represent the knee points (i.e., the lowpass cutoff frequencies at which the recognition performance reached 90% of the performance plateau) using the corresponding number of channels indicated on the abscissa. Other conventions as Fig. 1. Adapted from Xu and Zheng (2007) with permission from the Acoustical Society of America.

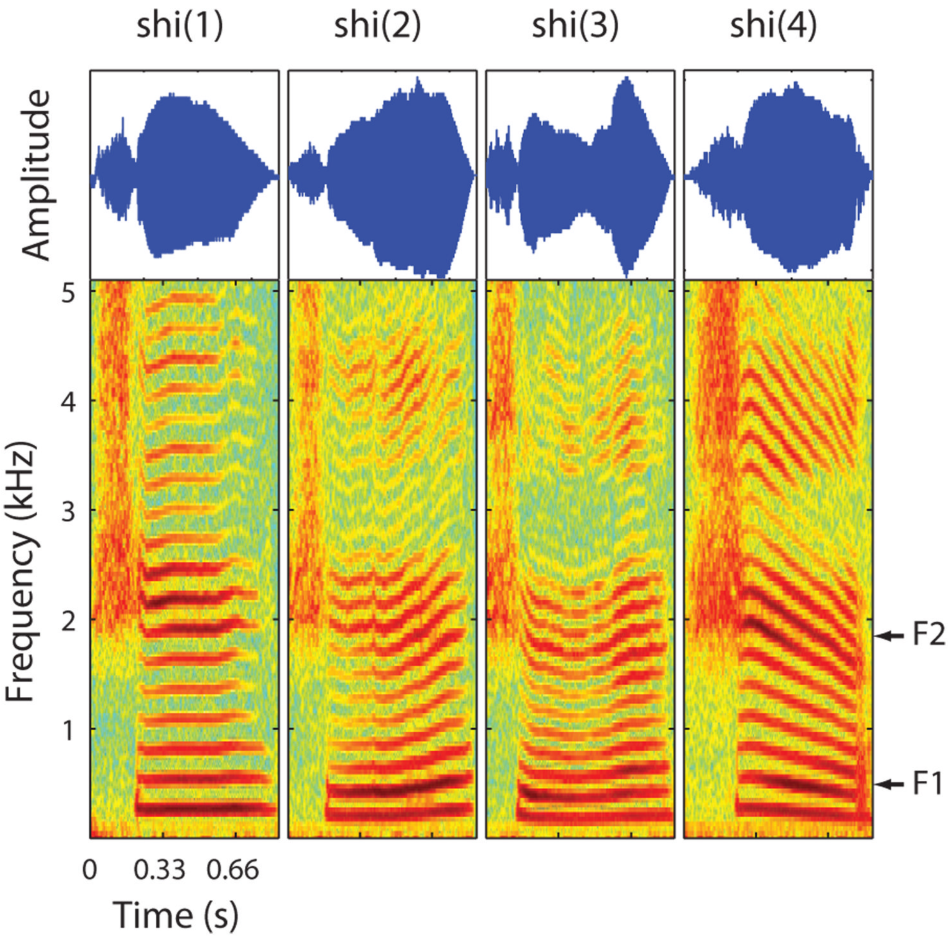

Fig. 3.

Time waveforms (top) and the narrowband spectrograms (bottom) of Mandarin Chinese syllable /shi/ spoken by a native Mandarin-speaking female adult. Panels from left to right show tone patterns 1 through 4. All tone tokens were of the same duration, 0.884 s. The arrows on the right indicate the first and second formants (F1 and F2) extracted in the middle of the vowel using the Praat software (Boersma and Weenink, 2007).

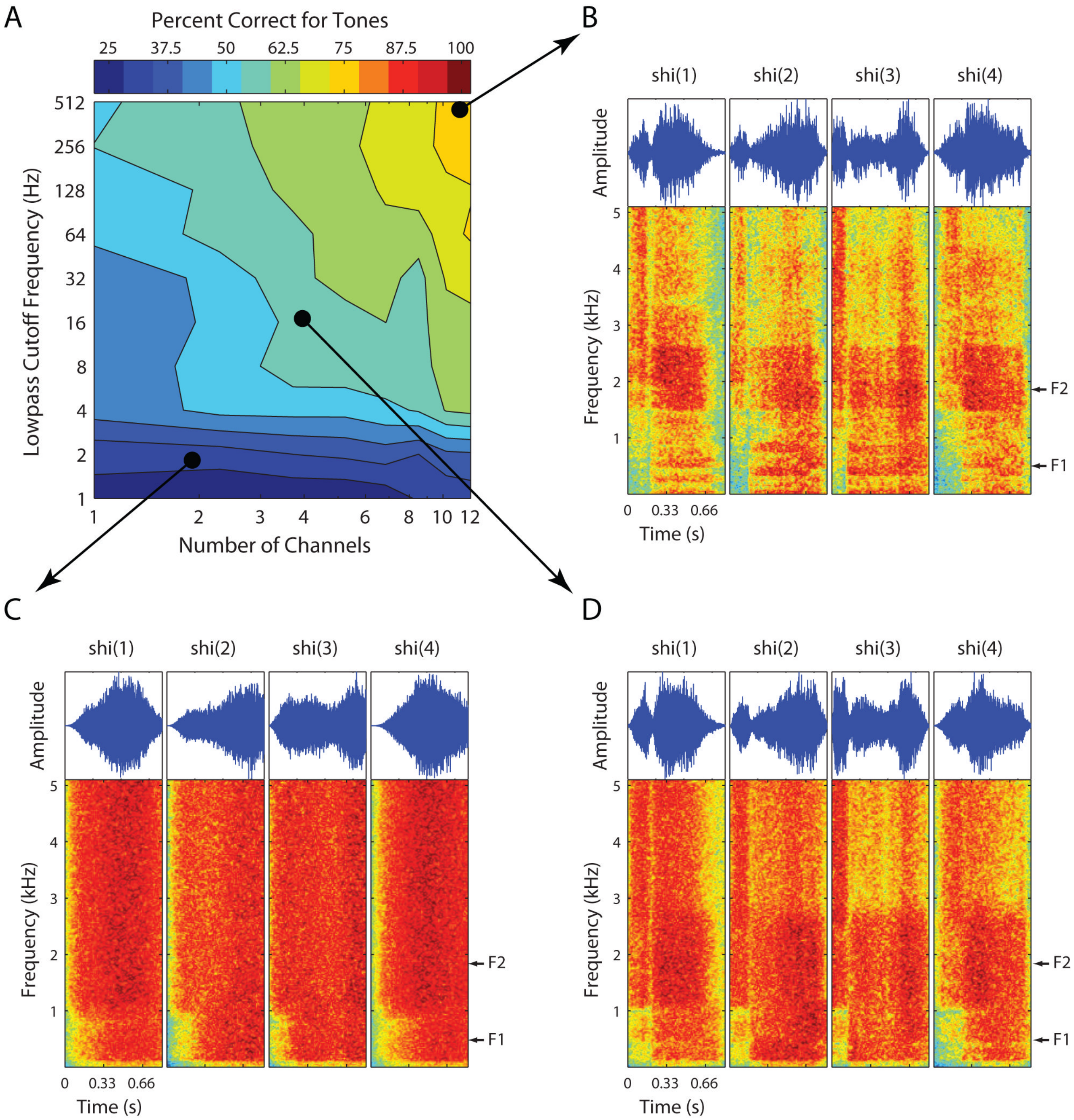

Fig. 4.

A: Mean tone recognition scores as a function of the number of channels and lowpass cutoff frequency. Other conventions as Fig. 1. Adapted from Xu et al. (2002) with permission from the Acoustical Society of America. B, C, and D: Time waveforms and the narrowband spectrograms of vocoder processed Mandarin Chinese syllable /shi/ in four tones shown in Fig. 3 with numbers of channels of 12, 2, and 4 and the lowpass cutoff frequencies of 512, 2, and 16 Hz, respectively. The short arrows on the right of each panel indicate the first and second formants (F1 and F2) extracted in the middle of the vowel of the original, unprocessed speech tokens shown in Fig. 3.