

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

**Developing Coding Schemes for Assessing Errors in Open-Set Speech
Recognition and Environmental Sound Identification¹**

Althea N. Bauernschmidt and Jeremy L. Loebach

*Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405*

¹ This research supported by NIH NIDCD R01 Research Grant DC00111 and NIH NIDCD T32 Training Grant DC00012 to Indiana University.

Developing Coding Schemes for Assessing Errors in Open-Set Speech Recognition and Environmental Sound Identification

Abstract. In the present study, we report on a series of coding schemes to classify errors in open-set recognition of speech and environmental sounds from previous perceptual learning experiments with cochlear implant simulations (Loebach and Pisoni, 2007; in press). Open-set responses to MRT and PB words were coded for place of articulation, manner of articulation and voicing errors in the identification of word initial and word final consonants. Open-set responses to meaningful and semantically anomalous Harvard sentences were coded for phonemic, lexical, and thematic errors in keywords. Open-set responses to environmental sounds were coded for errors in identifying the agent, action, and rhythm of the sounds. Overall, our coding scheme provided a more accurate assessment of performance producing higher percent correct recognition scores for the isolated words and environmental sounds than absolute coding schemes that simply identified entire words as correct or incorrect. Although time intensive, these coding schemes revealed perceptual elements with which subjects were having difficulty that were not apparent from the absolute coding scheme. The utility of open-set coding schemes is discussed for perceptual experiments with cochlear implant users, who often must make verbal responses to stimuli.

Introduction

According to the National Institute on Deafness and Other Communication Disorders, nearly 100,000 people worldwide have received cochlear implants as a treatment for profound hearing loss (National Institutes of Health, 2007). A cochlear implant is an auditory prosthesis that electrically stimulates the auditory nerve directly via electrodes placed in the cochlea. It thus bypasses damaged or missing hair cells that would normally stimulate the auditory nerve leading to the perception of sound. Although cochlear implant technology has been steadily improving over the past three decades, it does not and cannot restore normal hearing. The performance of the implant is influenced by several physiological and technological constraints and the resulting signal is severely spectrally degraded. However, considering the highly degraded signal that they receive, most cochlear implant users perform surprisingly well on speech recognition tasks in the quiet.

A common approach used to compensate for the degraded signal has focused on the optimization of acoustic speech information in the patterns of electric stimulation. The limited number of electrodes in the implant imposes fundamental limitations on the amount and type of information that can be transmitted. The number of electrodes that are available in the array is analogous to the number of channels of information that can be transmitted. In the normal hearing ear, there are approximately 20,000 channels transmitting detailed spectral and temporal information to the brain: in the best cochlear implant, however, there are only 24. One approach commonly used to determine how to best optimize the information in the speech signal has been to reduce the number of channels in the implant and determine the effect it has on speech perception. Generally, the more electrodes that are present in the cochlear implant, the more channels of information can be transmitted. A single channel implant only provides a crude signal that does not carry much fine spectral detail. Increasing the number of channels yields a richer signal and more spectral information. However, due to technological and physiological constraints, the number of electrodes that can be successfully implanted and utilized is limited.

It is possible to simulate the experience of hearing with a cochlear implant by using signal processing strategies similar those used in cochlear implants speech processors. This is done by band pass filtering the speech signal into the same number of channels as the number of electrodes in the array,

thereby limiting the number of channels of information that can be transmitted. The spectral detail is removed by replacing the spectral information in each band with a noise carrier, which is then modulated with the temporal envelope from the original band to simulate the temporal patterns of stimulation by the electrode array. Although there is no way to directly assess whether the simulations reproduce the actual experiences of a CI user, signal processing techniques based on CI speech processors may come close.

Studies using acoustic simulations of cochlear implants have demonstrated that the minimum number of channels that elicits high levels of speech perception can be as low as 4 (Shannon, Zeng, Wygonski & Ekelid, 1995) or as high as 20 (Dorman, Loizou, Fitzke & Tu, 1998) depending on the materials and task. These conflicting findings indicate that the type of information that is needed for good speech perception – whether purely temporal cues (Shannon et al., 1995), frequency cues (Dorman, Loizou & Rainey, 1997), or different listening strategies (Munson, Donaldson, Allen, Collison & Nelson, 2003) – is not known. Moreover, the acoustic environment, whether in quiet (Shannon et al., 1995) or in noise (Dorman et al., 1998), as well as the speech perception task that is being tested (Shannon et al., 1995; Dorman et al., 1997; Friesen, Shannon & Cruz, 2005) can influence the number of channels needed for good speech perception

Although most researchers have focused on the perception of speech under these degraded conditions, there are other important aspects of the world that are experienced through hearing. Another important and expected benefit of a cochlear implant is the improved recognition of environmental sounds. Many CI users report being able to perceive different types of sounds, such as “footsteps, slamming of doors, sounds of engines, ringing of the telephone, barking of dogs, whistling of the tea kettle, rustling of leaves, the sound of a light switch being turned on and off, and so on” (Food and Drug Administration, 2004). Moreover, the awareness of environmental sounds is often cited as an expected benefit from implantation (Clark, 2003). A particularly striking example of the expectations that implantees have for recognition of environmental sounds can be seen in the documentary *Sound and Fury*, a 2001 Academy Award Nominee for Best Documentary Feature. While the family of Heather, a 6 year old deaf child, is interviewing people about the benefits of their child receiving a cochlear implant the ability to perceive environmental sounds is constantly cited a reason for the surgery. Heather is told that she will be able to hear the birds chirping outside as well as many other environmental sounds. It is stressed that even if she won't be able to achieve full speech perception capabilities she will be able to hear other sounds (Aronson, 2000). Despite being a major motivation for cochlear implantation there has been little research into the perception of environmental sounds by individuals with cochlear implants, or normal hearing subjects listening to acoustic simulations of cochlear implants.

Moreover, few experiments have investigated the perception of environmental sounds in unprocessed auditory stimuli and much less is known about perception under conditions of auditory degradation. For normal hearing listeners, accuracy in open-set identification of 120 unprocessed environmental sounds exceeds 74% correct (Marcell, Borella, Green, Kerr & Rogers, 2000). For CI users, only closed-set experiments have been conducted, and accuracies as high as 79% have been reported when the stimulus set is limited to 37 sounds (Reed & Delhorne, 2005). Moreover, the important information for recognition of environmental sound stimuli has been elusive. While some studies have determined that temporal cues are important (Reed & Delhorne, 2005), others have found that the reliance on spectral or temporal information is conditional based on the specific type of sound (Gygi, Kidd, & Watson, 2004). Because only closed-set experiments on the recognition of environmental sounds have been conducted, it is unknown how well CI users perform in open-set tasks. A CI user's performance on an open-set task would provide information about which cues are important for the perception of environmental sounds, not only for CI users, but also for normal-hearing listeners as well. Performance on closed-set tasks may not be an accurate reflection of how well the CI is performing in relation to the perception of environmental sounds. Results on closed-set tasks may be inflated because the subject has learned the stimulus set or is relying primarily on contextual cues. These results do not allow

generalizations to how well CI users would perform on novel environmental sounds or sounds heard out of context.

By investigating how normal hearing listeners adapt to cochlear implant simulations, we can not only learn more about speech perception under degraded conditions, but may also discover information about perceptual learning that could have implications for rehabilitation of new CI users. This was the motivation for our previous perceptual learning experiment, which examined the effect of training on the recognition of speech and environmental sounds that were processed by an 8-channel sinewave vocoder (Loebach & Pisoni, 2007; in press). Using a diverse set of stimuli we compared whether training on words, sentences (meaningful or semantically anomalous), or environmental stimuli produced different levels of generalization to different materials using a pre-/post-test design. More importantly, we used an open-set task to assess how well subjects could identify the speech as well as environmental stimuli and determine what errors subjects make in doing so. Overall, we showed that training had a significant impact on generalization, with subjects who were trained on environmental stimuli performing as well on the speech perception tasks as subjects trained on speech. In addition, subjects who were trained on words did better when generalizing to new words and subjects who were trained on sentences did better when generalizing to new sentences.

One potential problem with our analysis is that we only scored each word as correct or incorrect, and did not take into account the types of errors subjects were making. Moreover, we did not describe the types of confusions subjects reported during the identification of environmental sounds. For the present study, we developed and implemented coding systems that provided a more detailed analysis of the errors that subjects made. Absolute coding schemes that score the number of keywords correct provide little insight into the errors that subjects make or why they make them. Understanding why subjects make mistakes in these perceptual learning tasks can aid in the evaluation of the strengths and weaknesses of different training paradigms and can lead to the development of better and more effective rehabilitation strategies for new CI users. In addition, knowing why subjects are making errors may provide insight into what information is successfully transmitted via the spectrally reduced signal, and how we can better optimize the perceptual cues available to the listener.

Method

Open-set responses to words, meaningful and anomalous sentences, and environmental sounds processed with an 8-channel sinewave vocoder were obtained from 125 normal hearing subjects (Loebach & Pisoni, 2007; in press). In that study, subjects were assigned to one of five training groups, and were explicitly trained to identify one type of stimulus materials, but were tested on all materials to assess generalization.

Coding Schemes for Single Words

Two of the sets of stimuli that we used in our original study were isolated monosyllabic words, varying in frequency of occurrence in American English and difficulty. The MRT word set is a corpus of 300 words made up of fifty lists of six rhymed variations on a common syllable (House, Williams, Hecker & Kryter, 1965). In each list of six rhymed words the word initial or word final consonant is systematically altered to produce six minimal pairs (e.g., ‘bat’, ‘bad’, ‘back’, ‘bass’, ‘ban’, ‘bath’). Ninety CVC words drawn from the MRT list were used in this experiment. The PB corpus consists of words whose phonemic composition approximates the statistical occurrence in American English (Egan, 1948). The corpus contains twenty lists of fifty monosyllabic words. Ninety unique words drawn from lists 1-3 of the PB corpus were used in this experiment.

To analyze the words, both PB and MRT, we divided the word into three parts: the word initial consonant or consonant cluster (C1), the word medial vowel (V), and the word final consonant or consonant cluster (C2). We developed a coding system to quantify the types of errors subjects made when listening to a degraded signal as well as how they adjust their perception after experience with the stimuli. Consonantal errors were classified by place of articulation (bilabial, alveolar or velar), manner of articulation (stop, fricative, or nasal), voicing (voiced or unvoiced), cluster insertion (adding a fricative or nasal to a plosive stop, or a plosive stop to a fricative or nasal), cluster deletion (deleting a fricative or a plosive stop), indeterminate (varying on more than two axes), or omissions (word was omitted completely). The response ‘plow’ for the target ‘cloud’ is an example of a word initial place error: the phoneme /k/ in ‘cloud’ is an unvoiced velar stop and the phoneme /p/ in ‘plow’ is an unvoiced bilabial stop but the two differ only in place of articulation. The response ‘bass’ for the target ‘mass’ is an example of a word initial manner error: /b/ is a voiced bilabial stop and /m/ is a voiced nasal bilabial stop and the two differ only in manner of articulation (stop versus nasal stop). The response ‘need’ for ‘neat’ is an example of a word final voicing error: /d/ and /t/ are both alveolar stops and differ only in voicing. An example of a cluster insertion error is the response ‘faint’ for ‘fate’: the word final consonant is the same in both except for the /n/ that was inserted before the /t/. An example of a cluster deletion error is the response ‘lush’ for ‘blush’: the /l/ in the word initial /bl/ sequence has been deleted, making a word initial cluster deletion error.

A response was considered indeterminate if there were more than two errors. For example, the response ‘out’ for the target word ‘earl’ was considered to be indeterminate at all places in the word, C1, V, and C2. In addition, a response could have been marked wrong as a combination of place, manner, or voicing. For example, if the word was ‘bat’ and the subject responded ‘tat’ then C1 would be wrong in both place and voicing, but not manner. Due to the response mode used in the original study (where subjects typed their responses on a computer keyboard) vowels were scored as correct or incorrect. If the word lacked a C1 or C2 consonant/consonant cluster (e.g., ‘earl’) the missing consonant/consonant cluster was considered correct if the subject correctly omitted it. When /r/ preceded a vowel, it was classified as part of C1 (ex: ‘drop’). When /r/ followed a vowel, it was classified as part of the vowel (ex: ‘earl’).

Coding Schemes for Sentences

Two different types of sentences were used in our original study: meaningful Harvard sentences, and semantically anomalous Harvard sentences. The meaningful Harvard sentences were drawn from the Harvard Sentence database (Karl & Pisoni, 1994). The database consists of seventy-two lists of ten meaningful sentences (IEEE, 1969). Sentences are phonetically balanced (relative to American English) and contain five keywords within a semantically rich meaningful sentence. Stimuli for the experiment consisted of twenty-five sentences taken from lists 1-10 of the Harvard Sentence database. The Anomalous Harvard sentences are semantically anomalous sentences that preserve the canonical syntactic structure of English, but contain thematically unrelated keywords. They were derived from the Harvard Sentence by taking the keywords from the 100 sentences in lists 11-20 and replacing them with words of equivalent semantic categories from lists 21-70 (Herman & Pisoni, 2000).

To identify the types of errors that were produced, whole keyword errors were classified as phonetic (errors made in a single phoneme or phoneme cluster), lexical (confusion with a lexically related word), thematic (filling in a word that is thematically related to the preceding or proceeding target), and indeterminate or omission. An error was considered a phonetic error if it deviated from the target word by one or two features (e.g., ‘bat’ for ‘back’), similar to the scoring of the MRT and PB words. A lexical error was the substitution of a word that is similar in meaning but not in sound (e.g., ‘boards’ for ‘planks’). A thematic error was a keyword error that was influenced by the surrounding words, or the overall interpretation of the sentence (e.g., ‘These days a chicken liver is a rare dish’ for ‘These days a chicken leg is a rare dish’).

Coding Schemes for Environmental Sounds

Environmental stimuli were drawn from the environmental signal database of Marcell and colleagues (2000). The database consists of stimuli recorded from a wide variety of acoustic environments developed for use in neuropsychological evaluation and confrontation naming studies (Marcell et al., 2000). Ninety stimuli from this database were used for the experiment. The environmental sounds were particularly challenging to score. As stated in the introduction, few experiments have investigated the perception of environmental sounds. As we were unable to find a conventional method of classifying environmental sounds that was comparable to the feature classification of speech sounds, we tried to identify the smallest number of subcategories that would differentiate the largest number of sounds.

Each of the environmental sounds was classified according to 3 features: the source or agent of the sound (the physical producer of the sound), the action that causes the sound, and the rhythmic or pitch information that differentiates it from other sounds. For the environmental sounds in the experiment, possible agents could include one or more of the following: animal (e.g., ‘wolf howl’), human (e.g., ‘child coughing’), wind (e.g., ‘boat horn’), glass (e.g., ‘glass shattering’), metal (e.g., ‘car crash’), wood (e.g., ‘door knocking’), liquid (e.g., ‘water boiling’), insect (e.g., ‘mosquito’), motor (e.g., ‘motorcycle’), plastic (e.g., ‘ping pong’) and string (e.g., ‘banjo’). The types of actions that could have caused the sounds presented were: burst (e.g., ‘harmonica’), strike (e.g. ‘basketball’), slide smooth (e.g., ‘sword fight’), slide rough (e.g., ‘violin’), tear (e.g., ‘paper tearing’), rumble (e.g., ‘thunder’), bubble (e.g., ‘water boiling’), blow (e.g., ‘whistling’), roll (e.g., ‘pinball’), crash (e.g., ‘glass shattering’), and buzz (e.g., ‘mosquito’). The types of rhythmic or pitch information that could distinguish the sounds presented were: pitch high (e.g., ‘baby crying’), pitch low (e.g., ‘boat horn’), pitch change (e.g., ‘airplane’), harmonic (e.g., ‘owl’), complex (e.g., ‘cash register’), transient (e.g., ‘rain’), periodic (e.g., ‘clock ticking’) and pulse (e.g., ‘gun shots’).

The difficulty in this method of scoring lies in scoring unique incorrect responses. For example, ‘cell phone’ was a common response for the sound ‘jackhammer’. Though we know that the sound of a jackhammer would have the agents metal and motor, the action of burst, and a periodic rhythm, we do not know along which dimension the sound of a cell phone differs from the sound of a jackhammer. Due to the high number of occurrences it was assumed that this was not a random error and that subjects were mistaking the sound of a cell phone set to vibrate for the sound of a jackhammer. Common responses like this were also classified according to our scheme and added to a large table for cross-reference. A portion of this table is reproduced in Appendix A for reference.

Results

Words

Analysis using the new individual phoneme based coding scheme increased performance on the MRT words across all five training groups ($M = 62.8\%$ correct) as compared to the absolute whole word correct coding scheme ($M = 34\%$ correct) used previously. When examined individually, performance was equivalent for word initial ($M = 58.6\%$) and word final ($M = 60.8\%$) consonants ($t(222) = -1.553$, $p = 0.123$) (Figure 1). For word initial consonants C1, subjects were in error 41.4 percent of the time. Of the responses that subjects could have made, 29.9% were errors in place of articulation, 17.2% were errors in manner of articulation and .3% were errors in voicing. Of the multiple feature errors, place and manner of articulation occurred 5.7% of the time, and errors on place and voicing, and manner and voicing were rare occurring less than .01% of the time. Cluster insertion and cluster deletion errors were also rare, occurring less than .01% of the time. For word final consonants C2, subjects were in error

39.1% of the time. Of the responses that could have been made, 19.3% were errors in place of articulation, 3.2% were errors in manner of articulation, and 1.5% were errors in voicing. Of the multiple feature errors, place and voicing errors occurred 3% of the time, place and manner errors occurred 2.5% and errors on manner and voicing were rare occurring less than .01% of the time. Cluster insertion errors occurred 5% of the time, while cluster deletion errors occurred less than .01% of the time. Comparing error prevalence across word initial and word final consonants revealed that subjects made significantly more place of articulation errors on C1 and C2 ($t(180) = 9.2356, p < 0.001$), but all other errors did not differ by consonantal position (all $p > 0.05$). Subjects made few errors in identifying vowels, scoring 69.1% correct.

Across the five training groups, training had a significant effect on the number of correct responses on MRT word initial ($F(4, 120) = 20.74, p < 0.001$) and MRT word final ($F(4, 120) = 6.59, p < 0.001$) consonants. Subjects scored significantly better on MRT C1 when trained on MRT words ($M = 73.5%$) than when trained on environmental sounds ($M = 56%$), Anomalous sentences ($M = 54.7%$), Harvard sentences ($M = 48.7%$), and PB words ($M = 60%$) (all $p < 0.001$). Subjects also scored significantly better on MRT C1 when trained on PB words ($M = 60%$) than when trained on Harvard sentences ($M = 48.7%$; $p < 0.001$). Subjects were also more accurate at identifying MRT C2 correct when trained on MRT stimuli ($M = 64.7%$) than when trained on Harvard sentences ($M = 54.8%$; $p < 0.001$), but equally well as when trained on PB words ($M = 65.2%$), Anomalous sentences ($M = 60.2%$) or Environmental sounds ($M = 59%$) (all $p > 0.05$).

Training had a significant effect on the number of place of articulation errors on C1 ($F(4, 120) = 19.618, p < 0.001$) and C2 ($F(4, 120) = 2.597, p = 0.041$). Subjects were significantly less likely to make place errors on C1 when trained on MRT words ($M = 17.2%$) than when trained on environmental sounds ($M = 31.7%$), anomalous sentences ($M = 34.2%$), Harvard sentences ($M = 38.5%$), and PB words ($M = 28%$) (all $p < 0.001$). Although there was a significant effect of training on C2 errors, no specific training stimuli affected the number of place errors made on word-final consonants. Training did have a significant effect on cluster insertions in word-final ($F(4, 120) = 7.292, p < 0.001$) but not word-initial ($F(4, 120) = 2.433, p = 0.051$) consonantal position. For word final consonants, training on MRT words ($M = 0%$) leading to significantly fewer cluster insertion errors than training on environmental sounds ($M = .8%$), Anomalous sentences ($M = 1.4%$), Harvard sentences ($M = .9%$), and PB words ($M = .8%$) (all $p < 0.03$). However, given the low occurrence of these errors across training groups this effect may be more apparent than real.

Analysis using the new individual phoneme-based coding scheme increased performance on the PB words across all five training groups ($M = 69.1%$ correct) as compared to the absolute whole word correct coding scheme ($M = 44%$ correct) used previously. When examined individually, subjects performed significantly better on word initial ($M = 70.2%$) than word final ($M = 66.04%$) consonants ($t(226) = 4.667, p < 0.001$) (Figure 1). For word initial consonants, subjects were in error 30.9 percent of the time. Of the errors that could have been made, 9.2% were errors in place of articulation, 1.9% were errors in manner of articulation and 2.3% were voicing errors. Of the errors that could be made on two features, place and voicing errors occurred 2% of the time, place and manner errors occurred 6.5% of the time, and manner and voicing errors were rare, occurring less than .01% of the time. Cluster insertion errors were also rare, occurring .5% of the time, while cluster deletion errors were more common, occurring 4.3% of the time. For C2, subjects were in error 30% of the time. Of the errors that could have been made, 12.1% were place of articulation errors, 2.0% were manner of articulation errors, and 2.2% were voicing errors. Of errors that could have been made on multiple features, .8% were place and voicing errors, 4% were place and manner errors, and manner and voicing errors were rare, occurring less than .01% of the time. Cluster insertion errors were more common, occurring 4.9% of the time, while cluster deletion errors were comparatively rare, occurring 1.8% of the time. Unlike the MRT words, subjects made significantly more place of articulation errors on the word final consonant ($M = 12.1%$)

than the word initial consonant ($M = 9.2\%$) ($t(248) = -5.98, p < 0.001$), but all other errors were equivalent across word initial and word final consonants (all $p > 0.05$). Subjects scored 71.1% correct on the PB vowels.

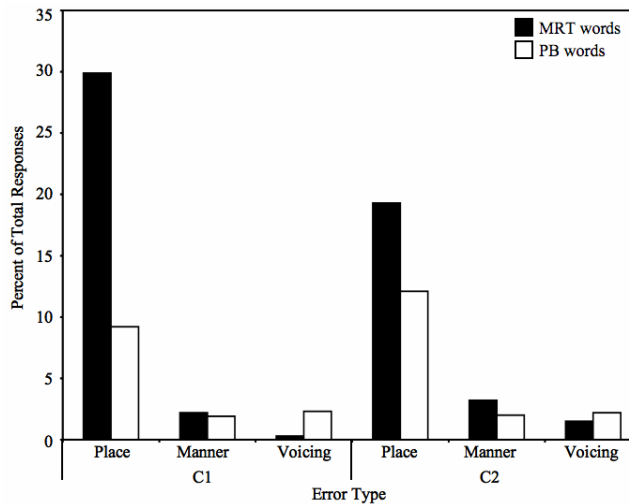


Figure 1. Comparison of errors made on MRT (black) and PB (white) word initial (C1) or word final (C2) consonants.

When comparing the performance on the PB words across the five training groups, training did not have a significant main effect on the percent correct recognition of C1 ($F(4, 120) = .78, p = 0.541$): all training conditions produced equivalent performance on C1. While there was a significant main effect of training on number of correct C2 responses ($F(4, 120) = 2.451, p = .05$), no specific training condition was significantly more likely to improve the subject’s word final PB consonant score (all $p > 0.074$). Training had significant effect on the number of voicing errors made on C1 ($F(4, 120) = 15.590, p < 0.001$) and C2 ($F(4, 120) = 2.805, p = 0.029$). Subjects were least likely to make voicing errors on C1 when trained on PB words ($M = .000$) than on any other type of material: MRT ($M = .3\%$), HS ($M = .2\%$), AS ($M = .3\%$), ENV ($M = .3\%$). Subjects were also significantly less likely to make voicing errors on C2 when trained on PB words ($M = .1\%$) than when trained on MRT words ($M = .2\%$), Anomalous sentences ($M = .3\%$) Harvard sentences ($M = .3\%$) and Environmental sounds ($M = .3\%$). However, given the low prevalence of errors, this effect may be more apparent than real.). Training also had a significant main effect on the number of manner and voicing errors that subjects made on both C1 ($F(4, 120) = 2.674, p = 0.035$) and C2 ($F(4, 120) = 3.578, p = 0.008$), but again, given the low prevalence of occurrence (less than 5% of the time) these differences may be more apparent than real.

Comparing only MRT and PB training groups, subjects performed equally well on the word initial consonants for PB ($M = 70.1\%$) and MRT words ($M = 72.5\%$; $t(48) = 1.598, p = 0.117$). For the word final consonant, however, subjects performed significantly better on PB words ($M = 70.0\%$) than MRT words ($M = 64.7\%$; $t(48) = 2.163, p = 0.036$). As reported above, training specificity had a greater effect for MRT consonants than PB consonants: MRT training produced significantly better performance on C1 for MRT words than training on other materials, whereas all forms of training were equally effective for C1 in PB words. Subjects also made significantly fewer place of articulation errors on C1 for PB words ($M = 10.0\%$) than on C1 for MRT words ($M = 17.1\%$; $t(36) = 4.58, p < .001$). Additionally, subjects made significantly fewer place of articulation errors on word final consonants for PB words ($M = 10.6\%$) than for MRT words ($M = 21.6\%$, $t(39) = .8652, p < 0.001$). The difference in error types as well as overall percent correct recognition scores on MRT and PB words may be due to the difference in the

phonemic composition of the two types of stimuli. The variability in the types of errors made on the PB words may be because that their phonemic composition approximates the statistical occurrence of those phonemes in American English. MRT words are not phonetically balanced relative to American English, and are composed of minimal pairs. This may predispose subjects to making specific types of errors (such as errors in place of articulation) since the stimuli can only differ on one or two dimensions. In addition to having more varied error types, PB words showed less training specificity than MRT words, which could also be due to the differences in phonetic balance.

Sentences

Across all five training groups, subjects scored 70.3% correct on Harvard sentence keywords. Of the responses, 14.2% were phonetic errors, 1.6% were thematic errors, and .4% were lexical errors (Figure 2). There was a significant effect of training on the number of Harvard sentence keywords correctly identified ($F(4, 120) = 3.47, p = .01$): subjects' performance improved when trained on Harvard sentences ($M = 76.5%$) than on MRT ($M = 68%$) or PB ($M = 68%$) words (both $p < 0.03$). Subjects trained on Anomalous sentences ($M = 69%$) and Environmental sounds ($M = 69.8%$), however, performed as well as subjects trained on Harvard sentences (both $p > 0.07$). Subjects made significantly less phonetic errors when trained on Harvard sentences ($M = 8%$) than on MRT words ($M = 18%$), Environmental sounds ($M = 16%$), Anomalous sentences ($M = 16%$), or PB words ($M = 14%$) ($F(4, 120) = 13.8, p < .001$). Subjects also made significantly fewer thematic errors ($F(4, 120) = 5.161, p = .001$) when trained on Harvard sentences ($M = .8%$) than on Environmental sounds ($M = 2%$) Anomalous sentences ($M = 2%$), MRT words ($M = 1.6%$) and PB words ($M = 1.3%$). However, given the low prevalence of thematic errors overall, this effect may be more apparent than real.

Across all five training groups, subjects scored 55% correct on Anomalous sentence keywords. Of the total Anomalous keyword responses, 29.5% were phonetic errors, .6% were thematic errors, and there were no lexical errors (Figure 2). There was a significant effect of training on number of Anomalous sentence keywords correct ($F(4, 120) = 21.05, p < .001$), and subjects performed better on Anomalous sentences when trained on Anomalous sentences ($M = 68%$) than on MRT words ($M = 50%$), Environmental sounds ($M = 50%$), Harvard sentences ($M = 60%$), and PB words ($M = 47%$). Subjects made significantly fewer phonetic errors ($F(4, 120) = 17.4, p < .001$) when trained on Anomalous sentences ($M = 21%$) than when trained on MRT words ($M = 33%$), Environmental sounds ($M = 34%$), Harvard sentences ($M = 28%$), or PB words ($M = 29.8%$).

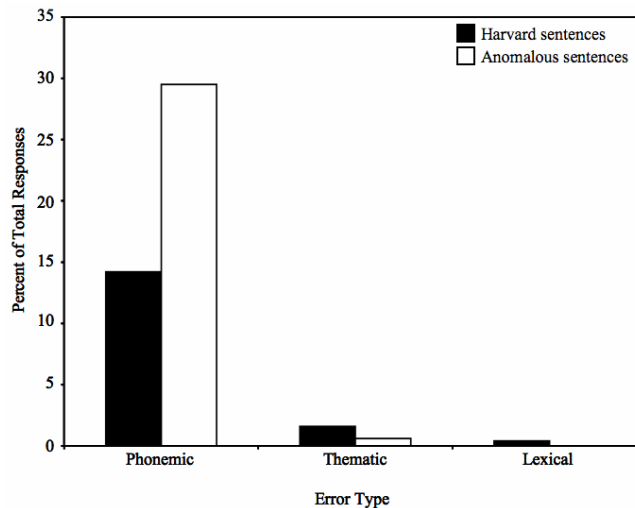


Figure 2. Comparison of errors made in Harvard (black) and Anomalous sentences (white).

Comparing the two training groups, subjects performed significantly better on Harvard sentences than Anomalous sentences ($t(42.9) = -3.36, p = .002$). Subjects made significantly more phonetic errors on Anomalous sentences than on Harvard sentences ($t(48) = 9.39, p < .001$). Subjects also made significantly more lexical errors when trained on Harvard sentences than when trained on Anomalous sentences ($t(33.6) = -2.08, p = .045$). There were no significant differences in thematic errors across the two training groups. Because the Anomalous sentences are derived from the Harvard sentences, any difference in performance between the two is presumably due to sentence context. It is therefore interesting that the only error type not to be significantly affected by training is thematic errors. If subjects were filling in misheard items with thematically related items, it would suggest that they are using a contextually driven strategy for keyword identification. Since the thematic errors were rare overall, and did not differ between the two training groups, it would suggest that subjects were primarily using an acoustic phonetic perceptual strategy for identification. Moreover, no differences in performance were observed between the new coding scheme and the previous coding scheme (Loebach & Pisoni, 2007; in press) due to both coded for whole keywords only. The new coding scheme, however, did provide an additional level upon which to assess subject performance and determine the cognitive strategies that subjects employ when listening to these sentences.

Preliminary Results for Environmental Sounds

Error coding for the environmental sound recognition tasks is ongoing due to the complexity of the coding scheme and the large number of novel responses that subjects report. Coding is complete for one of the five training groups (MRT training group), and these preliminary data are presented below.

Using the new coding scheme, a significant increase in performance was observed for environmental sounds, increasing from 37.6% correct for the absolute coding scheme to 49.6% correct for the 3-tiered coding scheme. For the Agent valence, subjects were correct 46.9% of the time. Of the errors that subjects made, 17.3% were determinate (i.e., could be classified under the new coding scheme) and 38.8% were indeterminable, for a total error rate of 53.13%. Table 1 displays an error matrix of possible responses (columns) to target agents (rows). Cell values represent the total percentage of responses made. Values along the diagonal (grey shaded cells) indicate the percentage of incorrect responses that were

	Animal	Human	Insect	Liquid	Wind	Glass	Metal	Wood	Paper	Rubber	String	Motor	Unk.
Animal	.09	.01	.03	0	.02	0	.13	0	0	0	0	.01	.43
Human	.12	0	0	.02	.01	0	.09	.01	0	0	.01	0	.37
Insect	.04	0	.19	.02	0	0	.02	0	0	0	0	0	.21
Liquid	0	0	0	.01	.01	0	.04	0	0	0	0	0	.27
Wind	.01	.01	0	.01	0	0	.10	.01	0	0	.01	.07	.40
Glass	0	0	0	0	0	0	.44	0	0	0	0	0	.56
Metal	.01	0	0	0	.01	0	.01	0	0	0	0	0	.30
Wood	0	0	0	.01	.01	.01	0	.02	0	0	0	0	.30
Paper	.06	0	0	.06	0	0	.24	0	0	0	0	0	.59
Rubber	0	0	0	.07	.03	0	.17	0	0	0	0	.03	.63
String	0	0	0	0	.09	0	.17	0	0	0	.02	0	.66
Motor	0	0	0	0	.02	0	.03	0	0	0	0	.10	.17

Table 1. Error matrix displaying the frequency of responses for each agent. Target agents appear in column 1, and response agents appear in row 1.

within the same agent class as the target (e.g., responding “duck” to a stimulus of a “cow”). Row sums indicate the total percentage of incorrect responses, and can be subtracted from one to obtain the percent

correct recognition scores for each target agent. Of the possible agents that subjects described in their responses, Metal was the most prevalent agent error, with subjects indicating metallic agents 53% of the time. The next most prevalent error was Animal, with subjects indicating animal agents 12% of the time.

For the Action valence, subjects correctly identified the appropriate action 52.1% of the time. Of the errors that subjects made, 8.4% were determinate (i.e., could be classified under the new coding scheme) and 39.5% were indeterminable (for a total error rate of 47.9%). Of the possible actions that subjects described in their responses (Table 2), Strike was the most prevalent action error, with subjects indicating striking actions 52% of the time. The next most prevalent error was Rumble, with subjects indicating rumbling actions 18% of the time. A comparison of the frequency of incorrect actions being selected revealed that there was a significant main effect of action ($F(10,709) = 14.299, p < 0.001$). Post hoc Bonferroni tests revealed that this effect was driven entirely by the Strike action (with striking actions being described significantly more often than any other action; all $p < 0.001$). No other action was selected significantly more often than any other.

	Blow	Bubble	Burst	Buzz	Crash	Roll	Rumble	Slide R	Slide S	Strike	Tear	Unk.
Blow	0	0	.02	0	.01	0	.02	0	0	.08	0	.47
Bubble	0	0	.08	0	0	0	.04	0	0	0	0	.58
Burst	.01	0	0	0	.02	0	0	0	0	.11	0	.51
Buzz	0	0	0	0	0	0	0	0	0	0	0	.30
Crash	0	0	0	0	0	0	0	0	0	.02	0	.23
Roll	0	0	0	0	0	0	0	0	0	.08	0	.83
Rumble	.01	0	0	0	.01	0	0	.02	0	.01	0	.27
Slide R	0	0	0	0	0	0	.07	0	.02	0	0	.34
Slide S	.02	0	0	0	.01	0	0	0	0	.04	0	.41
Strike	0	0	0	0	0	0	0	0	0	.01	0	.26
Tear	0	0	0	0	0	0	0	0	0	.06	0	.82

Table 2. Error matrix displaying the frequency of responses for each action. Target actions appear in column 1, and response actions appear in row 1.

For the Rhythm valence, subjects correctly identified the appropriate rhythm 49.9% of the time. 6.7% of errors were determinate (i.e., could be classified under the new coding scheme) and 43.5% were indeterminable (for a total error rate of 50.1%). Of the possible rhythms that subjects described in their responses (Table 3), Pitch High or Low was the most prevalent rhythm error, with subjects indicating high or low pitches actions 42% of the time. The next most prevalent error was Periodic, with subjects indicating periodic rhythms 25% of the time. A comparison of the frequency of incorrect rhythms being selected revealed that there was a significant main effect of rhythm ($F(6,413) = 2.94, p = 0.008$). Post hoc Bonferroni tests failed to differentiate rhythms, and given the small amount of determinate errors, the main effect of rhythm may be more apparent than real.

	Complex	Harmonic	Periodic	Pitch C	Pitch H/L	Pulse	Transient	Unk.
Complex	0	.01	.03	.02	.02	0	0	.60
Harmonic	0	0	.04	0	.02	0	0	.56
Periodic	0	0	0	.02	.02	0	0	.47
Pitch C	0	0	0	0	0	.02	0	.24
Pitch H/L	0	.01	0	0	.03	0	0	.46
Pulse	0	.01	.02	.05	.01	0	0	.38
Transient	.01	.01	.06	.01	.03	.01	.01	.46

Table 3. Error matrix displaying the frequency of responses for each rhythm. Target rhythms appear in column 1, and response rhythms appear in row 1.

Discussion

Using the new coding schemes, several interesting findings emerged. For single words (MRT and PB), subjects' performance increased using the new 3-tier coding scheme over the absolute whole word correct coding scheme used previously. Performance was comparable on word initial and word final consonants for both MRT and PB words, indicating that subjects were equally likely to make errors in C1 and C2. Place of articulation errors were most common regardless of word type (MRT or PB) or consonantal position (word initial or word final). This is most likely due to the reduced spectral detail in the sinewave vocoded stimuli. These findings are similar to those of Shannon and colleagues (1995), who found that place errors were the most common for vocoded stimuli regardless of the number of channels used in synthesis or low pass filter cutoff frequencies used to derive the amplitude envelopes. The finding that subjects made few voicing errors overall is not surprising because temporal information, which provides many cues for voicing, was preserved through the use of the 400 Hz filter for envelope detection. This result is also similar to those of Shannon and colleagues (1995) who found that varying the amount of temporal information by using higher and lower cutoff frequencies for amplitude envelope detection altered the number of voicing errors made. When higher frequency filter cutoffs were used (preserving more temporal information) voicing errors were less prevalent than when lower frequency filter cutoffs were used (preserving less temporal information and increasing the number of voicing errors).

The difference between the performance on the two word sets and the types of errors that were made on each is interesting as well. Subjects made more varied errors (i.e., more errors of different types) on PB words despite performing better overall (69% correct) than on MRT words (63%). Subjects also showed less training specificity on PB words, which could be due to differences in phonetic balance between MRT (not phonetically balanced) and PB words (phonetically balanced relative to American English). The MRT words were composed of minimal pairs, varying along only one or two dimensions, and are therefore limited to words that have five other words that rhyme with them. By comparison, testing on PB words is probably a better assessment of an individual's open-set word recognition under degraded conditions, whereas testing on MRT words may provide a better assessment of feature discrimination and reception.

That subjects performed better on the contextually rich Harvard sentences (70.3% correct) than on the semantically anomalous sentences (55% correct) is unsurprising because subjects could use semantic context to make informed guesses if they were unsure of identity of the keyword. The coding scheme revealed that the majority of the keyword errors were phonetic errors and very few were lexical or thematic errors as was expected. The lack of thematic errors is particularly intriguing, since both the Harvard and the Anomalous sentences are grammatically licit. We predicted that subjects would make errors based on the semantic context of the surrounding words in the sentences, but this did not appear to be the case. Previous work with SPIN sentences has demonstrated that word predictability significantly influences sentence recognition (Kalikow, Stevens & Elliott, 1977). This predictability, however, was limited to the thematic relationship of the final word in the sentences to the preceding stem, not the thematic relationship between each word in the sentence relative to one another. Thus, it could be the case that when the target item is isolated in the sentence, predictability may constrain the response set, but when all words are thematically unrelated, predictability may not be invoked as a perceptual strategy. Further research will be necessary to determine the extent of the relationship between predictability and sentence semantic structure. In addition to there being few thematic errors overall, there were no differences between in thematic error prevalence between meaningful and anomalous sentences. The high number of phonetic errors on sentences as well as the large number of place errors on single word stimuli suggests that it would be effective to train people on phonetic contrasts when adapting to degraded stimuli.

Using the new 3-tier coding scheme, subjects' performance was much higher on environmental sounds than using the absolute scheme used previously. We initially designed this 3-tier coding scheme to provide more information about the cues that are important for the perception of environmental sounds. However, due to the high number of indeterminate errors, it is unclear how useful the present coding scheme will be to that end. The agent, or object or event that produced the sound, was the only valence that could be differentiated reliably under the new coding scheme (with 17% of the errors made being determinate errors). Both action and rhythm errors could only be reliably differentiated 8 and 6 percent of the time respectively. This is somewhat disappointing, since differentiating errors in rhythm was one of the reasons that we designed this coding scheme. Future work will re-examine the three valences to determine whether more powerful and reliable coding schemes can be devised to differentiate subject errors.

Investigating how normal hearing listeners adapt to cochlear implant simulations not only provides more information about speech perception under degraded conditions, but may have implications for rehabilitation strategies for new CI users. Moreover, many experiments with CI users utilize open-set recognition of stimuli where the subject listen to a target stimulus and verbally report what they perceived. The development of the coding strategies presented here will therefore be extremely useful to open-set studies with CI users. In an ongoing experiment in our lab, CI users are being tested on the same materials as reported on here so that a comparison can be made between the performance of normal hearing subjects on CI simulations and CI users themselves. The same coding strategies implemented here will be used to assess the types of errors that CI users make on these same stimuli to determine if CI users and normal hearing process these materials in similar ways. We hope that such a detailed error analysis will reveal differences in perceptual processing strategies that are used by the two groups of subjects, which would have important implications for training and rehabilitation paradigms for postlingually deafened cochlear implant users.

References

- Aronson, J. (Producer and Director)(2000). *Sound and Fury* [Motion Picture]. United States: Aronson Films.
- Clark, G.M. (2003). "Rehabilitation and Habilitation." In *Cochlear Implants: Fundamentals and Applications*. pp 654-706 New York NY: Springer-Verlag.
- Dorman, M.F., Loizou, P.C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *Journal of the Acoustical Society of America*, *102*, 2403-2411.
- Dorman, M.F., Loizou, P.C., Fitzke, J. & Tu, Z. (1998). The recognition of sentences in noise by normal hearing listeners using simulations of cochlear implant signal processors with 6-20 channels. *Journal of the Acoustical Society of America*, *104*(6), 3583-3585.
- Egan, J.P. (1948). Articulation testing methods. *Laryngoscope*, *58*, 955-991.
- Food and Drug Administration. (2004). Cochlear Implants: Frequently Asked Questions. Retrieved August 9, 2007, from <http://www.fda.gov/cdrh/cochlear/faq.html>.
- Friesen, L.M., Shannon, R.V., & Cruz, R.J. (2005). Effects of stimulation rate on speech recognition with cochlear implants. *Audiology and Neuro-otology*, *10*, 169-184.
- Gygi, B., Kidd, R.R., & Watson, C.S. (2004). Spectral-temporal factors in the identification of environmental sounds. *Journal of the Acoustical Society of America*, *115*(3), 1252-1265.
- Herman, R., & Pisoni, D.B. (2000). Perception of elliptical speech by an adult hearing-impaired listener with a cochlear implant: some preliminary findings on coarse-coding in speech perception. In *Research on Spoken Language Processing Progress Report No 24* pp 87-112 Bloomington IN: Speech Research Laboratory, Indiana University.

- House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, 37, 158-66.
- IEEE. (1969). IEEE recommended practice for speech quality measurements. IEEE No 297 New York: Author.
- Kalikow, D.N., Stevens, K.N., & Elliot, L.L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America*, 61, 1337-1351.
- Karl, J.R. & Pisoni, D.B. (1994). Effects of stimulus variability on recall of spoken sentences: A first report. In *Research on Spoken Language Processing Progress Report No 19*, pp 145-193 Bloomington IN: Speech Research Laboratory, Indiana University.
- Loebach, J.L., Bent, T., Peterson, N., Hay-McCutcheon, M. & Pisoni, D.B. (2008). The perception of speech and environmental sounds by normal hearing listeners and cochlear implant users. Poster to be presented at the 31st Annual Midwinter Meeting of the Association for Research in Otolaryngology, Phoenix, AZ.
- Loebach, J.L. & Pisoni, D.B. (in press). Perceptual learning of spectrally degraded speech and environmental signals. *Journal of the Acoustical Society of America*.
- Loebach, J.L & Pisoni, D.B. (2007). Perceptual learning under a cochlear implant simulation. In *Research on Spoken Language Processing Progress Report No. 28*. Bloomington, IN: Speech Research Laboratory, Indiana University.
- Marcell, M.M., Borella, D., Greene, M., Kerr, E. & Rogers, S. (2000). Confrontation naming of environmental sounds. *Journal of Clinical and Experimental Neuropsychology*, 22(6), 830-864.
- Munson, B.G.S., Donaldson, G.S., Allen, S.L., Collison, E.A., & Nelson, D.A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *Journal of the Acoustical Society of America*, 113(2), 925-935.
- National Institutes of Health (2007). "Cochlear Implants." Retrieved August 1, 2007, from <http://www.nidcd.nih.gov/health/hearing/coch.asp>.
- Reed, C.M., & Delhorne, L.A. (2005). Reception of environmental sounds through cochlear implants. *Ear and Hearing*, 26, 1 48-61.
- Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303-304.

Appendix A: Excerpted coding decisions for the environmental sound coding scheme

	'Airplane'		'Baby crying'		'Banjo'		'Basketball'	
Airplane'	Mo/Ru/PC	Mo/Ru/PC	Mo/Ru/PC	Hu/B/PH	Mo/Ru/PC	St/P/Cx	Mo/Ru/PC	R/S/Pe
'Baby crying'	Hu/B/PH	Mo/Ru/PC	Hu/B/PH	Hu/B/PH	Hu/B/PH	St/P/Cx	Hu/B/PH	R/S/Pe
Agent	St/P/Cx	Mo/Ru/PC	Rhythm/Chitch	Hu/B/PH	St/P/Cx	St/P/Cx	St/P/Cx	R/S/Pe
'Banjo' Animal	B	Burst	Complex	Hu/B/PH	R/S/Pe	St/P/Cx	R/S/Pe	R/S/Pe
'Basketball'	R/S/Pe	Blow	Harmonic	Hu/B/PH	Hu/Ru/Pu	St/P/Cx	Hu/Ru/Pu	R/S/Pe
B	Hu/Ru/Pu	Bubble	Pitch Change	Hu/B/PH	A/BI/H	St/P/Cx	A/BI/H	R/S/Pe
Glass	A/BI/H	Crash	Periodic	Hu/B/PH	Me,PL/SR,C/T	St/P/Cx	Me,PL/SR,C/T	R/S/Pe
Belch	Hu/Ru/Pu	Cluck	Sharp Hit	Hu/B/PH	Wi/BI/PI,Pu	St/P/Cx	Wi/BI/PI,Pu	R/S/Pe
Birds'	A/BI/H	Roll	Rich Low	Hu/B/PH	Wi/S/Pe	St/P/Cx	Wi/S/Pe	R/S/Pe
Insect	Cr	Rumble	Pulse	Hu/B/PH	Wo/R,Cr/T	St/P/Cx	Wo/R,Cr/T	R/S/Pe
'Blinds closing'	Me,PL/SR,C/T	Strike	Transient	Hu/B/PH	Mo/PI/Ru,Cr/PH	St/P/Cx	Mo/PI/Ru,Cr/PH	R/S/Pe
Boat horn	Wi/BI/PI,Pu	Slide Rough	Me/B/PH	Hu/B/PH	Me/B/PH	St/P/Cx	Me/B/PH	R/S/Pe
Engines	Wi/S/Pe	Slide Smooth	Ru,Me/SS,Cr/Cx	Hu/B/PH	Ru,Me/SS,Cr/Cx	St/P/Cx	Ru,Me/SS,Cr/Cx	R/S/Pe
Motor	Ru	Tear	Wi/BI/Pu	Hu/B/PH	Wi/BI/Pu	St/P/Cx	Wi/BI/Pu	R/S/Pe
Bowling	Wo/R,Cr/T	Buzz	Me/S,SR/Cx	Hu/B/PH	Me/S,SR/Cx	St/P/Cx	Me/S,SR/Cx	R/S/Pe
Paper	Mo/PI/Ru,Cr/PH			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Camera	Mo/PI/Ru,Cr/PH			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Plastic	Me/B/PH			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Can opening	Me/B/PH			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Hubber	Ru,Me/SS,Cr/Cx			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Car crash	Ru,Me/SS,Cr/Cx			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
String	Wi/BI/Pu			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Car horn	Wi/BI/Pu			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Wind	Me/S,SR/Cx			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
'Wash register'	Me/S,SR/Cx			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Wood	A/BI,Ru/PC			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
'Chickens'	A/B/Cx			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
'Child coughing'	Hu/B/Pe			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe
Church bell'	Me/S/PL			Hu/B/PH	A/BI,Ru/PC	St/P/Cx	A/BI,Ru/PC	R/S/Pe

Abbreviations used in the above table

