# Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions[a)]

Valerie Hazan[b)] and Rachel Baker
*Speech, Hearing, and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1E 1PF, United Kingdom*

(Received 11 June 2010; revised 13 July 2011; accepted 18 July 2011)

This study investigated whether speech produced in spontaneous interactions when addressing a talker experiencing actual challenging conditions differs in acoustic-phonetic characteristics from speech produced (a) with communicative intent under more ideal conditions and (b) without communicative intent under imaginary challenging conditions (read, clear speech). It also investigated whether acoustic-phonetic modifications made to counteract the effects of a challenging listening condition are tailored to the condition under which communication occurs. Forty talkers were recorded in pairs while engaged in "spot the difference" picture tasks in good and challenging conditions. In the challenging conditions, one talker heard the other (1) via a three-channel noise vocoder (VOC); (2) with simultaneous babble noise (BABBLE). Read, clear speech showed more extreme changes in median F0, F0 range, and speaking rate than speech produced to counter the effects of a challenging listening condition. In the VOC condition, where F0 and intensity enhancements are unlikely to aid intelligibility, talkers did not change their F0 median and range; mean energy and vowel F1 increased less than in the BABBLE condition. This suggests that speech production is listener-focused, and that talkers modulate their speech according to their interlocutors' needs, even when not directly experiencing the challenging listening condition. © *2011 Acoustical Society of America.*

[DOI: 10.1121/1.3623753]

PACS number(s): 43.70.Mn, 43.70.Fq, 43.70.Jt [CYE] Pages: 2139–2152

## I. INTRODUCTION

Even though the acoustic-characteristics of speech are to a great extent determined by physiological factors such as vocal tract size and vocal fold length, talkers still have a degree of control over the acoustic-characteristics of the speech that they produce (e.g., Johnson and Mullennix, 1997). This control can be used to modify speech to meet the needs of listeners, as can be seen in speaking styles such as child-directed (e.g., Fernald and Kuhl, 1987; Burnham *et al.*, 2002) or foreigner-directed speech (e.g., Uther *et al.*, 2007; Van Engen *et al.*, 2010) as well as speech to listeners in adverse listening conditions. In this study, we investigate to what degree spontaneous speech produced with communicative intent to counter intelligibility-challenging conditions differs from speech produced for communication purposes under more ideal conditions and from speech produced without communicative intent under imaginary challenging conditions (i.e., when talkers are asked to read sentences clearly). We also investigate whether the acoustic-phonetic modifications made by talkers are attuned to the specific challenging condition that their interlocutors are experiencing. This study therefore investigates whether speech production is guided by interlocutors' communicative needs.

The Hyper-Hypo (H&H) theory of speech production (Lindblom, 1990) is a useful framework for our study as it discusses how the control that talkers have over their speech production is used to maximize communication efficiency in different communicative situations. According to the H&H theory, during speech communication, there is an ongoing tension between the talker's desire to minimize articulatory effort (i.e., by producing hypo-articulated speech) and the need for effective communication; phonetic variability occurs to deal with this tension as talkers can produce a range of articulations on a hypo- to hyper-speech continuum. Hypo-articulated speech, which demands the least degree of effort on the part of the talker, is adequate when there is a significant degree of signal-independent linguistic-contextual information present. Hyper-articulated speech is typically produced in response to listeners' increased difficulty in understanding speech, which is either due to impoverished language knowledge by the listener (i.e., if the listener is a child or a second-language speaker), to the presence of a communication barrier in the form of an adverse listening condition (e.g., background noise, other voices) or situations in which linguistic-contextual information is not sufficient to convey the message (e.g., transmission of flight coordinates by air traffic controllers).

The production of hyper-articulated or clear speech is therefore seen as integral to the communicative process between two or more talkers. It is perhaps contradictory,

[a)] Portions of this work were presented in "Acoustic-phonetic characteristics of naturally elicited clear speech in British English," a poster presented at the 157th Meeting of the Acoustical Society of America, Portland, OR, 18-22 May 2009; "Acoustic characteristics of clear speech produced in response to three different adverse listening conditions," a poster presented at the Psycholinguistic Approaches to Speech Recognition in Adverse Conditions Workshop, Bristol, 8-10 March 2010; and "Spot the different speaking styles: Is 'elicited' clear speech reflective of clear speech produced with communicative intent?" a poster presented at the British Association of Academic Phoneticians (BAAP) Colloquium 2010, London, 29-31 March 2010.

[b)] Author to whom correspondence should be addressed. Electronic mail: v.hazan@ucl.ac.uk

0001-4966/2011/130(4)/2139/14/$30.00

therefore, that the most commonly used methodological approach in clear speech studies does not involve any communicative intent. Indeed, the usual methodology has been to record talkers while reading sentences or words "normally" and then to ask them to repeat the task while speaking "as if to a deaf person" or similar instruction (e.g., Picheny *et al*., 1985, 1986; Bradlow *et al*., 2003; Kain *et al*., 2008). A body of recent research has investigated which acoustic-phonetic features are enhanced when talkers produce a clear speaking style by reading sentences clearly under instruction (for recent reviews, see Smiljanić and Bradlow, 2009; Uchanski, 2005). Talkers typically produce quite consistent hyper-articulation within a recording session, and the acoustic-phonetic characteristics of this type of clear speech have also been shown to be relatively consistent across studies. These include reductions in speaking rate (Picheny *et al*., 1986; Smiljanić and Bradlow, 2005; Uchanski *et al*., 1996), higher energy in the mid-frequency region of the frequency spectrum (Krause and Braida, 2004), higher mean fundamental frequency (Picheny *et al*., 1986) and fundamental frequency range (Picheny *et al*., 1986), longer and more frequent pauses (Picheny *et al*., 1986; Liu and Zeng, 2006), and more expanded vowel spaces (Picheny *et al*., 1986; Moon and Lindblom, 1994; Ferguson and Kewley-Port, 2002, 2007; Bradlow *et al*., 2003; Krause and Braida, 2004; Bradlow, 2002). Some studies have also found that final consonants are more likely to be released (Picheny *et al*., 1986) and found increased modulation in the intensity envelope (Krause and Braida, 2004). In intelligibility tests using these kinds of materials, clear speech has been found to be of benefit to many groups of listeners including non-native talkers (Bradlow and Bent, 2002), learning-impaired talkers (Bradlow *et al*., 2003) and deaf talkers (Liu *et al*., 2004).

A question that has received relatively little attention is the degree to which this type of clear speech varies from speech produced in spontaneous speech when trying to overcome the effects of a challenging listening environment. These two types of speech differ in two key aspects: (a) presence vs absence of communicative intent and (b) imagined or real challenging conditions. The impact of communicative intent was addressed in a study of Lombard speech (i.e., the specific speaking style produced by talkers while directly subjected to noisy conditions) in which three talkers carried out a map description alone vs a task similar to the Map Task (Anderson *et al*., 1991) with another talker (Garnier *et al*., 2010). Certain parameters that were enhanced in the non-communicative task (i.e., increases in vocal intensity and related parameters) were amplified further in the communicative task; further modifications occurred in the communicative task that were not present in the non-communicative task, for certain talkers at least. Cooke and Lu (2010) using a Sudoku-solving task, also found significant acoustic-phonetic differences in the Lombard speech produced when the speaker was describing the problem-solving process aloud vs when solving the task with a partner. These studies point to the need to investigate specific speaking styles, such as clear speech, in communicative settings, such as in problem-solving tasks between two talkers in good or challenging listening conditions. In our study, we compare speech produced with communicative intent (in a problem-solving task between two people in good and challenging conditions) with read speech produced normally and when talkers are instructed to speak clearly. Our expectation is that clear, read speech elicited through instruction will be more consistently hyper-articulated than spontaneous speech produced to counteract a challenging listening environment. A further expectation is that presence or absence of communicative intent will also impact on speech characteristics: read "conversational" speech will be more "enhanced" than spontaneous speech produced in interaction.

A further objective of our study was to investigate whether the acoustic-phonetic characteristics of speech are specifically adapted to the challenging condition that it is seeking to overcome. We know that speaking styles targeted at specific populations of listeners, such as children or non-native speakers, share some characteristics with "clear speech" but also differ in certain respects. For example, infant-directed speech (IDS) and "clear speech" both have higher mean fundamental frequency and wider fundamental frequency range, a more expanded vowel space, and slower speech rate than conversational speech (e.g., Fernald and Kuhl, 1987; Kuhl *et al*., 1997; Burnham *et al*., 2002). However, there are also important differences, and there have been arguments as to whether IDS can be considered as a form of hyper-speech (Fernald, 2000; Davis and Lindblom, 2001). IDS and foreigner-directed speech (FDS) share some common features (e.g., expanded vowel space), but other characteristics, such as heightened pitch, occur in IDS but not FDS (Uther *et al*., 2007). Another speaking style, pet-directed speech, also shared features with IDS, such as heightened pitch, but did not lead to an expanded vowel space (Burnham *et al*., 2002). Certain aspects of these speaking styles (such as pitch range) may therefore have an affective function, while others, such as vowel space expansion, have a linguistic function. A study comparing vowels produced by the same group of talkers in IDS, Lombard speech, hyper-speech, and citation speech suggested that talkers may manipulate acoustic-phonetic features of vowels according to the perceived needs of the listener, e.g., their degree of linguistic competence, environmental experience, knowledge of discourse context (Wassink *et al*., 2006). All these studies therefore show some evidence of modulation of the acoustic-phonetic characteristics of the clear speech to the specific needs of different types of listeners. In our study, we investigate whether such modulation occurs even though the talker is not directly experiencing the challenging listening condition affecting their interlocutor: Do we modulate our speech differently when speaking with someone with a cochlear implant or someone communicating via telephone in a noisy room?

To move from laboratory speech to speech produced with communicative intent in good and challenging conditions, a technique is needed to elicit spontaneous speech that is controlled to the extent that comparable materials are produced when communication between two talkers is either easy or difficult. The Map Task (Anderson *et al*., 1991), a cooperative problem-solving task in which an "instruction giver" has to communicate details of a trajectory on a map to an "instruction follower," is such a technique and has been used in a number of studies to elicit spontaneous dialogs with a constrained topic and specific keywords. Recently, Bradlow and colleagues (Van Engen *et al.*, 2010) developed the "diapix" task, also a problem-solving task involving two talkers. The two talkers are each presented with a picture and have to collaborate to find a number of differences between the two pictures. Specific keywords can be elicited via the differences that need to be found; as in the Map Task, the content of the speech produced is quite similar across talkers and conditions as the topic of the conversation is constrained. The diapix task shares many similarities with the Map Task (Anderson *et al.*, 1991) but involves a more balanced contribution between the two talkers and is likely to be more varied in terms of the sentence structures produced than the Map Task, which involves mostly quite short commands (Baker and Hazan, 2010). Using the diapix task, the communicative situation can be controlled to naturally elicit clear speech in one or both talkers. For example, in Van Engen *et al.* (2010), communication difficulty was controlled by comparing the speech produced by a participant when communicating with a talker with a shared native language or with a nonnative talker. Simple measures, such as the time taken to complete the task or the balance of speech between the two talkers, were sensitive enough to show differences in communication difficulty between native talker pairs, native-nonnative pairs, and nonnative pairs with matched or unmatched L1s.

In our study, we use the diapix task to compare the acoustic-phonetic characteristics of speech that was naturally elicited in two challenging communicative conditions that differ in the type of degradation they entail. Many studies have investigated the speech characteristics of talkers directly affected by a challenging listening environment (e.g., Cooke and Lu, 2010). Our study differs in that the communication impairment is placed on one talker only (the "adverse listening" or AL talker), and we are investigating how the talker communicating with this "impaired" listener but who is hearing normally (the "normally hearing" or NH talker) clarifies his or her speech to maintain effective communication. As NH talkers are not experiencing the degraded speech, they have to adapt their speech purely in response to the covert or overt feedback received from the AL talkers, possibly using their own knowledge of the adverse listening condition.

Two "communication barrier" conditions were chosen that differ in the type of degradation they impose on the speech; these conditions simulate that of a hearing person who is communicating with a cochlear implant user (VOC condition) or with a person who is in a noisy environment (BABBLE condition). In the VOC condition, the AL talker hears the NH talker's speech passed through a three-channel noise-excited vocoder, which spectrally degrades the signal and removes much of the pitch information. In the BABBLE condition, the AL talker hears the NH talker's speech in background multi-talker babble noise, which has a masking effect on the signal.

Different predictions can be made as to which acoustic enhancements made by the NH talker would or would not be helpful in each of these conditions. For the BABBLE condition, we can get some sense of what clarifications might help overcome the effects of babble noise from studies of the changes that talkers make to their voice when directly subjected to noise (e.g., Lombard speech) (Lane and Tranel, 1971). With the same eight-talker babble noise as used in our study, Lu and Cooke (2008) found that talkers reading sentences while directly affected by noise showed changes in sentence duration, root mean square energy, mean F0, spectral center of gravity, and voiced/unvoiced ratio. Vowel F1 frequency also increased significantly. Higher energy is clearly likely to be beneficial in increasing intelligibility due to the masking effects of the noise, and a slowing down of the speaking rate gives the listener greater opportunity to glimpse acoustic information in the babble noise (Lu and Cooke, 2008). An increased spectral center of gravity and general shifting of energy to higher frequency regions may also decrease the masking effects of the noise. In perception studies, Lombard speech provided intelligibility gains compared to normal speech (Lu and Cooke, 2008). This suggests that "efficient" enhancements made when speaking to an interlocutor who is hearing speech with background babble noise may show similar characteristics to Lombard speech itself. The VOC condition entails a spectral degradation of the speech passed through the vocoder and also a loss of much of the pitch and voicing information. To our knowledge, there are no studies of the acoustic-phonetic characteristics of "adult CI-directed" speech that may give hints to the adjustments that are made in real-life communication with adults with cochlear-implants, although there have been studies examining the communication strategies used by CI users when communicating with their hearing peers (e.g., Tye-Murray, 1995; Ibertsson *et al.*, 2009). In the absence of direct evidence, we can consider what changes the talker whose voice is being vocoded could most usefully make to compensate for this degradation. We hypothesize that changes to F0 median and range would have no positive benefit to the AL talker as pitch information is mostly lost in the vocoded signal; boosting mid-frequency energy is also unlikely to clarify the signal as the intelligibility problems are due to poor frequency resolution rather than poor audibility. The most helpful clarifications are expected to be a slowing down of the speaking rate and expansion of the vowel range (at least for certain vowel contrasts with a similar F1/F2 structure).

Our acoustic-phonetic analyses of the NH talker's spontaneous speech therefore focused on whether there was evidence of different levels of acoustic-phonetic adaptation in the VOC and BABBLE conditions. Many acoustic-phonetic measures have been used in the investigation of clear speech as detailed in the preceding text. Here, measures were selected that addressed the differences expected between the two conditions and that could reliably be measured from unconstrained spontaneous speech. These included measures expected to vary between the VOC and BABBLE conditions: two pitch measures (median pitch and pitch range) and a measure reflecting average intensity in the 1–3 kHz frequency region containing many key acoustic cues. Other measures were included that were not expected to vary: speech rate and measures reflecting vowel hyper-articulation (F1 and F2 range).

In addition to the comparison across VOC and BABBLE conditions, the degree to which speech produced in these intelligibility-challenging conditions differs from clear speech elicited via instruction was investigated by comparing, for the same group of talkers, the acoustic-phonetic characteristics of the speech produced in the "no barrier," VOC, and BABBLE dialog situations with speech produced when instructed to read a set of sentences normally and clearly.

## II. METHOD

### A. Participants

Forty native speakers of Southern British English (20 male; 20 female) aged between 19 and 29 yr served as main participants in the study, i.e., participants whose speech was analyzed. They were all students or staff from University College London in the UK. Participants volunteered with a friend of the same gender so there were 20 "friend" pairs (10 male pairs, 10 female pairs). Participants read accent-revealing sentences before being accepted onto the study to ensure that they were from a homogeneous accent group (Southern British English). Another eight native speakers of Southern British English (4 male; 4 female) who fitted the preceding criteria were recruited as confederates for the recording session involving background noise (BABBLE condition).

All participants were screened for normal hearing thresholds (20 dB hearing level or better for the range 250–8000 Hz) and reported no history of speech or language disorders. All but two of the main participants had no specific experience of communicating with people with speech and language difficulties. Participants were not aware of the purpose of the recordings. They were paid for their participation and were debriefed afterward.

### B. Materials

#### *1. Spontaneous speech task*

For the dialog conditions, we used the diapixUK task (Baker and Hazan, 2010), which is an extension (in terms of the number of picture pairs available) of the diapix task created by Bradlow and colleagues (Van Engen *et al.*, 2010). It is an interactive "spot the difference" game for two people that allows for recordings of natural spontaneous speech. Each diapixUK task consists of two versions of the same cartoon picture that contain 12 differences. Each person is given a different version of the picture and is seated in a separate sound-treated room (without a view of the other person). The pair communicates via headsets to locate the 12 differences between the two pictures. The experimenter monitored the recording from outside both of the recording rooms.

Twelve pairs of pictures were created for this study and form the diapixUK materials. The pictures included hand-drawn scenes produced by an artist that were then colored in; these were designed to be fairly humorous to maintain interest in the task (see Fig. 1 for an example of one of the picture pairs). Each picture included different "mini-scenes" in the four quadrants of the picture, and the differences were fairly evenly distributed across the four quadrants. These could be differences in an object or action across the two pictures (e.g., green ball in picture 1 vs red ball in picture 2; holding the ball in picture 1 vs kicking the ball in picture 2) or omissions in one of the pictures (e.g., missing object on a table in one picture). In each picture pair, each difference was designed to encourage elicitation of 1 of 36 keywords. Each keyword is a monosyllabic CV(C) word that belongs to a (near) minimal word pair with the /p/-/b/ or /s/-/ʃ/ contrasts in initial position (e.g., *pear*/*bear*; *sign*/*shine*). This allows for the analysis of the production of these two contrasts in different speaking styles, although the analyses presented here are focused on more general acoustic-phonetic measures. The 12 picture pairs belong to one of three themes: beach scenes, farm scenes, and street scenes with four picture pairs per theme. The keyword set was divided into three, and each set of 12 keywords was used for a different picture set. As a result, completion of three diapix tasks (1 beach, 1 street, and 1 farm scene) would be likely to result in the production of the whole set of 36 keywords.

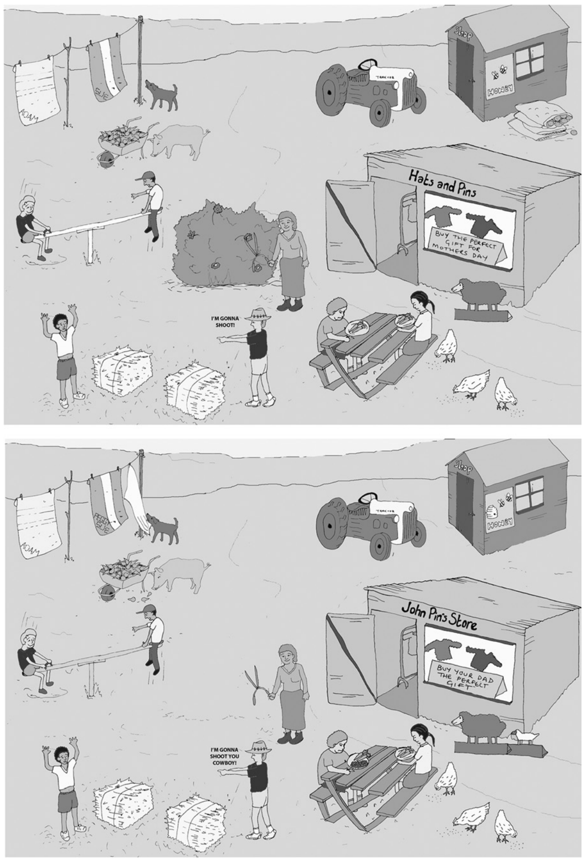
Hats and Pins
BUY THE PERFECT GIFT FOR MOTHERS DAY
I'M GONNA SHOOT!
John Pin's Store
BUY YOUR DAD THE PERFECT GIFT
I'M GONNA SHOOT YOU COWBOY!

FIG. 1. A black and white version of a pair of diapixUK pictures that are part of the "farm" theme. Twelve differences have to be found between the two pictures.

A pilot study verified that all picture pairs were of equal difficulty by comparing the average number of differences found per picture within a set time for eight pairs of pilot participants. The pilot study also ascertained that the learning effect of participating in more than one picture task was minimal. A training picture pair was also developed and contains 12 differences that are not related to the keyword set.

#### *2. Read speech task*

A set of 144 sentences was recorded by each participant. This included four sentence pairs for each of the 18 /p-b/, /s-ʃ/

keyword pairs. Within each sentence pair, keywords were matched for prosodic position and preceding phonetic context/phoneme. Keyword position in the sentence was varied between pairs. Example sentences are given in the following text:

The old lady ate the *peach*
The young children loved the *beach*

For the recording session, all sentences were randomized and presented on a screen one at a time. The keywords were not italicized in the sentences presented in the recording session.

## C. Procedure

Each participant took part in five recording sessions on separate days: the first three sessions involved diapix recordings with another talker, while the remaining two involved the recording of read materials individually (see a graphical representation of the overall test design in Fig. 2). Beyerdynamic DT297PV headsets fitted with a condenser cardioid microphone were used in all recording sessions, and the speech was recorded at a sampling rate of 44 100 Hz (16 bit) using an EMU 0404 USB audio interface and Adobe AUDITION (diapix sessions) or DMDX (read sentence sessions) software. For the diapix tasks, two-channel recordings were made with the speech of each talker on a separate channel to facilitate the transcription and acoustic analysis stages.

### *1. Diapix sessions (3 sessions)*

All participants took part in three sessions involving diapix recordings. Each participant did the first two sessions with the same friend. In the third session, half of the participants carried out the diapix tasks with one of the English confederates and half with one of the non-native confederates.[2] All participants were presented with each diapixUK picture pair once only. The order in which pictures were completed was counterbalanced across conditions and participant pairs following a modified Latin square design. In all sessions, participants were told to start the task in the top left corner of the picture and work in a clockwise manner around the scene. For each task, the experimenter stopped the recording either once the 12 differences were found or when the participants could not locate all differences after at least 15 min had lapsed.

*a. Session 1.* This session was completed with both talkers hearing each other in good listening conditions ("no barrier" condition—NB). Both participants were asked to contribute to finding the differences to encourage a natural and balanced conversation between the two participants. Each of the three recordings lasted around 8 min on average, so on average, 25 min of speech recordings were obtained per pair across the three pictures, which gave around 7.8 min of speech per person for the NB condition, once pauses, silences, and non-speech portions (laughter, etc) had been excluded.

*b. Sessions 2 and 3.* All participants completed the diapix task in two adverse conditions. All 40 participants did the task in the VOC condition. Half of the participants then did the task in the BABBLE condition with a native confederate. The confederates[1] were the talkers who were hearing their partner under adverse listening conditions (AL talker). The level of impairment was such that the participant hearing normally (NH talker) needed to speak clearly to communicate successfully with their partner; as noted in the preceding text, it is the NH talker's speech that is of interest in this study. For all communication-barrier conditions, the NH participant was encouraged to take the lead in the conversation. This was to discourage the AL talker, i.e., the person who was in an adverse listening situation but whose speech was being heard normally by the other participant, from dominating the conversation to alleviate communication difficulty.

In the VOC condition, the AL talker heard the speech of the NH participant after it had been processed in real-time through a three-channel noise-excited vocoder, which was the three-channel version of the vocoder described in Rosen *et al.* (1999). This has the effect of significantly spectrally degrading the speech, as the speech spectrum was processed

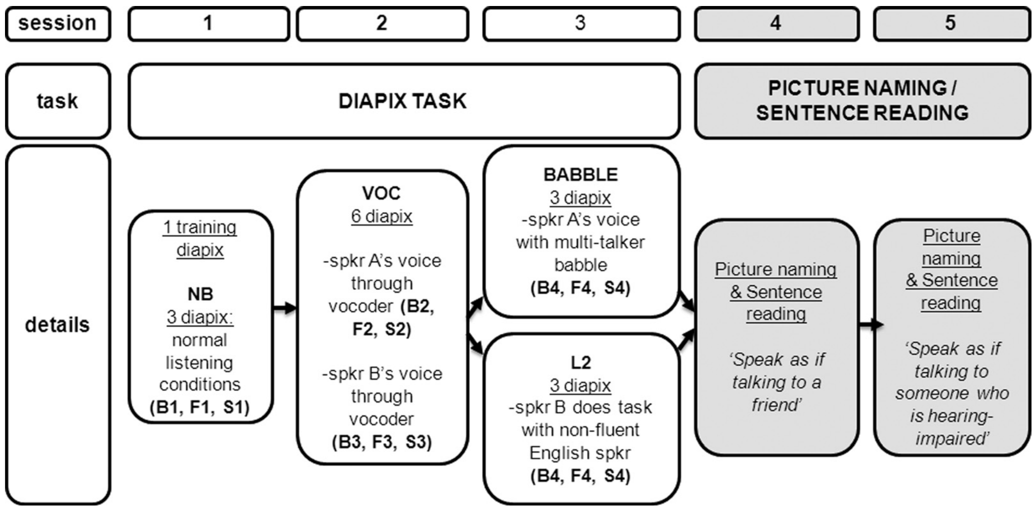
session
1
2
3
4
5
task
DIAPIX TASK
PICTURE NAMING / SENTENCE READING
details
1 training diapix
NB
3 diapix: normal listening conditions
(B1, F1, S1)
VOC
6 diapix
-spkr A's voice through vocoder (B2, F2, S2)
-spkr B's voice through vocoder
(B3, F3, S3)
BABBLE
3 diapix
-spkr A's voice with multi-talker babble
(B4, F4, S4)
L2
3 diapix
-spkr B does task with non-fluent English spkr
(B4, F4, S4)
Picture naming & Sentence reading
'Speak as if talking to a friend'
Picture naming & Sentence reading
'Speak as if talking to someone who is hearing-impaired'

FIG. 2. Diagram showing the order of presentation of the different experimental conditions over the five sessions. In the diapix sessions, "B," "F," and "S" denote beach, farm, and street scenes, respectively. There were four different diapix picture sets for each of these three themes.

through three filters only. A three-channel vocoder introduced enough difficulty to the task to necessitate the NH participant to clarify their speech while still allowing enough communication to do the task. As there is a significant learning effect when listening to vocoded speech (Davis *et al.*, 2005; Bent *et al.*, 2009), immediately prior to the diapix tasks in this session, each participant completed a 10-min vocoder-familiarization task where they listened to a story presented in three-channel vocoded speech and, after each sentence, clicked on the words that they heard. Feedback was given after each trial to reinforce familiarization to the distorted speech. Bent *et al.* (2009) found asymptote in the adaptation to an eight-channel sinewave vocoder after the presentation of around 60 meaningful sentences, which suggests that a significant part of the learning process was likely to be accounted for within the familiarization period although this may be slower here due to the use of a three-channel rather than eight-channel vocoder. This training meant that the NH talker had a sense of the adverse listening condition that their interlocutor was experiencing in the diapix task. Each pair of talkers completed six diapix tasks in total: three when the first participant's speech was vocoded and three when the other participant's speech was vocoded. Across the three pictures, on average 28.8 min of recordings were obtained per pair, giving on average 12 min of speech for the NH talker whose speech was being analyzed in the VOC condition.

In the BABBLE condition, 20 of the participants (11 male, 9 female) did three diapix tasks with a native confederate of the same gender, who was the AL talker. The speech of the normal-hearing participant was mixed with eight-talker babble (Lu and Cooke, 2008) before being channeled through to the confederate's headphones, at an approximate level of 0 dB SNR. The confederate had previously done the training diapix task in normal listening conditions as means of familiarization with the task procedure. The NH talker was told that their interlocutor would hear their speech in a background of lots of voices mixed together, which would be quite loud compared to their voice. Across the three pictures, on average 28.3 min of recordings were obtained per pair, giving about 12 min of speech for the NH talker whose speech was being analyzed in the BABBLE condition.

#### 3. *Read sentences sessions (2 sessions)*

In Session 4, participants were presented with sentences on a computer screen and were asked to read them "casually as if talking to a friend." There were 144 sentences presented in a pseudo-randomized order in 12 blocks of 12 sentences with a short break between blocks. In Session 5, they did exactly the same task but were instructed to read the sentences "clearly as if talking to someone who is hearing impaired." In each session, each participant also completed a picture naming task. Participants were presented with pictures representing the 36 diapixUK keywords in a random order and were required to say the name of the picture in one of two sentence frames: "I can see a <noun keyword>" or "the verb is to <verb keyword>." These data are currently being used to investigate the relationship between phoneme category dispersion and intelligibility and are not reported here.

In summary, the London UCL Clear Speech in Interaction Database (LUCID) corpus includes read, conversational and read, clear materials and spontaneous speech dialogs in good and intelligibility-challenging conditions for 40 talkers from a homogeneous accent group with a total of 110 h of recordings.[3]

### D. Data processing

For all diapix files, each channel containing the speech of one of the participants (excluding confederates) was orthographically transcribed using freeware transcription software from Northwestern University's Linguistics Department (WAVESCROLLER) to a set of transcription guidelines based on those used by Van Engen *et al.* (2010). The transcripts were automatically word-aligned to the sound files using NUALIGNER software, also from Northwestern, which created a PRAAT TextGrid (Boersma and Weenink, 2001, 2010). The word-level alignment was hand-checked in approximately two-thirds of the file set. All speech files were normalized to an average amplitude of 15 dB (with soft limiting) in Adobe AUDITION. For the read files, the transcriptions of the sentences were also word-aligned to the sound files as in the preceding text.

The acoustic-phonetic measures carried out on the spontaneous and read speech recordings included measures of fundamental frequency median and range, mean word duration (reflecting speech rate), mean energy in the 1–3 kHz range of the long-term average spectrum of speech, and vowel space.

#### 1. *Fundamental frequency: median and range*

Fundamental frequency analyses were done in PRAAT on each of the recordings for the NH talkers for each picture task in each condition using a time step of 150 value/s. For each individual diapix recording, a PRAAT script was used to calculate the median fundamental frequency (using the "meanst" function in PRAAT) and interquartile range (i.e., the difference between the values calculated using the "quant1st" and "quant3st" functions). The F0 measures were calculated in semitones relative to 1 Hz. The measures were averaged over the three picture tasks to obtain median F0 and interquartile range values in semitones (re 1 Hz) per talker per condition. A median value was preferred to the mean to reduce the effect of inaccurate period calculations, which are likely in spontaneous speech, while semitones were used to facilitate comparisons across male and female talkers.

#### 2. *LTAS measure*

Long-term average spectrum (LTAS) was also measured via a PRAAT script, based on the use of the "Ltas" function in PRAAT (with the bandwidth set at 50 Hz). Separate measures were obtained for each picture task in each condition for each of the 40 talkers. The PRAAT script was used to carry out the following operations on the single-channel speech recordings after they had been normalized for peak intensity.

First, silent portions were removed using the silence annotations within the PRAAT TextGrid; then the LTAS was calculated using a 50 Hz bandwidth, and the values for the first 100 bins (covering a 0–5000 Hz bandwidth) were obtained. A 1–3 kHz mean energy value (ME1-3kHz) was calculated as the mean of the bin values between these two frequencies.

### 3. *Word duration measure*

Mean word duration (MWD) was used as a measure reflecting the average speaking rate of a talker in a given condition. To obtain MWD, a PRAAT script was first used to calculate the duration of each of the orthographically annotated regions of the speech recording. These were then imported into a spreadsheet, and each annotated region was tagged as one of the following: agreement (AGR), breath (BR), filler (FIL), "garbage" (GA), hesitation (HES), laughter (LG), silence (SIL), and speech (SP). The GA label was used for regions containing sounds that were not produced by the talker, such as microphone pops and background noise; the AGR label marked agreements such as "okay," "yeah," etc. MWD was calculated by dividing the total duration of SP regions by the number of words produced in the recording. Again, MWD was initially calculated per picture and then averaged over the three pictures to get a measure of mean word duration per talker per condition.

### 4. *Vowel measures*

A number of steps were needed to obtain measures of vowel F1 and F2 range from the spontaneous speech recordings. First, a PRAAT script was run to remove annotations for all except content words (i.e., function words, unfinished words, hesitations, fillers, etc). Then, an SFS program (Huckvale, 2008) was used to obtain a phonemic transcription of the content words in the file and to carry out a phoneme-level alignment to the speech waveform. Formant estimates were then obtained in SFS for each vowel segment. Median vowel formant values were calculated for all monophthongs in content words per talker per condition. Even though errors are possible at several stages of this analysis (phoneme transcription and alignment, formant estimations) when spontaneous speech is being analyzed, the amount of speech on which the vowel estimates are based, and the use of median rather than mean values would mitigate the effect of these errors, especially for the point vowels used for the F1/F2 range calculations, which were typically numerous.[4] The range values were based, for each talker, on the difference between the lowest and highest median F1 and F2 values across the vowel range.

## III. RESULTS

### A. In the diapix tasks, is there evidence that the communication barrier conditions were successful in modifying the speech produced by the NH talker (relative to the NB condition)?

The NB condition and two communication barrier conditions (VOC, BABBLE) were compared to ascertain that our communication barrier conditions were successful in making communication more effortful between the two talkers. As a measure of transaction difficulty, the time taken to find the first eight differences in the pictures was calculated (not all pairs managed to find all 12 differences by the maximum allotted time, but all had found at least 8 of the differences).[5] This measure of task completion time discriminated across native and nonnative talker groups in Van Engen *et al.* (2010). The data are shown in Table I. A repeated-measure analysis of variance (ANOVA) revealed that transaction time was significantly longer for the VOC than for the NB condition [F(1,37) = 66.4; $P < 0.001$]. There was a condition by picture order interaction [F(2,74) = 6.8; $P < 0.005$]: Transaction time for the VOC condition got shorter with practice as might be expected due to the learning effect when listening to vocoded speech, suggesting that not all learning had been completed by the end of the vocoder-familiarization task. For the 20 talkers who carried out the BABBLE condition, transaction time for this condition was also significantly longer than for the NB condition [F(1,19) = 5.97; $P < 0.05$], but no other effects or interactions were significant. There was therefore no indication of a significant learning effect across the three picture tasks. Both communication-barrier conditions therefore led to longer transaction times than the NB condition.

TABLE I. Mean time in seconds taken for talker pairs to find the first eight differences for each of the pictures in the NB ("no barrier"), VOC, and BABBLE conditions. Standard deviation measures are given in italics. Three pictures were presented per condition.

| | NB (N = 38) | | VOC (N = 38) | | BABBLE (N = 20) | |
|---|---|---|---|---|---|---|
| | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| Picture 1 | 266 | *95* | 366 | *97* | 338 | *122* |
| Picture 2 | 244 | *73* | 326 | *82* | 303 | *87* |
| Picture 3 | 262 | *83* | 303 | *84* | 330 | *109* |
| Mean | 257 | *74* | 331 | *81* | 324 | *93* |

Longer transaction times in the communication barrier conditions than the NB condition are likely to indicate greater task difficulty and thus an increased need for the NH talker to clarify their speech. To investigate whether the communication barrier conditions did indeed result in the NH talker modifying their speech characteristics, a perceptual rating experiment was run to determine whether listeners perceived the speech produced by the NH talkers in the VOC and BABBLE conditions as clearer than speech produced by the same talkers in the NB condition. For every talker, two short samples of speech were excised from each of their three conversations in the NB, VOC, and BABBLE conditions, resulting in six speech samples per condition per talker. The samples were excised from as close as possible to the 10th and 20th turns in each conversation after a number of criteria had been met: they had to be between 2 and 3 s long, were either a whole intonational phrase or the end of a phrase and did not occur after a miscommunication. The samples were therefore chosen according to objective criteria rather than for their distinctiveness or speaking style and specifically excluded speech expected to be hyper-articulated due to a recent miscomprehension. Thirty-six native southern British English talkers with normal hearing were the participants in this rating experiment. The randomized samples were presented to listeners over headphones across two sessions, and listeners rated the clarity of each sample using a 7-point scale (1, very clear, to 7, not very clear).

First, a repeated-measures ANOVA was conducted on the ratings data for the NB and VOC snippets, which were taken from the full set of talkers (20 male, 20 female). Mean ratings per talker were lower in the VOC condition (2.5) than in the NB condition (3.4) [F(1,35) = 113.5, $P < 0.001$], suggesting that listeners judged the speech from the VOC condition as clearer than the speech from the NB condition. The mean ratings for the BABBLE condition were obtained for the 20 talkers recorded in that condition (10 male, 10 female). The mean ratings per talker were lower in the BABBLE (2.4) condition than the NB condition (3.3) [F(1, 35) = 68.7, $P < 0.001$]. These data show that random speech samples from the VOC and BABBLE conditions were perceived as clearer than speech samples taken from the NB condition. This suggests that the NH talkers were indeed using strategies to clarify their speech in response to their interlocutors' adverse listening conditions.

### B. How stable are the acoustic-phonetic measures within-condition given that they are based on spontaneous speech?

Prior to investigating the effect of communication barriers on the acoustic-phonetic characteristics of conversational speech, it is important to ascertain the stability of these measures within-condition. Indeed, as we are measuring spontaneous speech, the lexical content of the speech varied across different picture tasks for a given condition, and this variation in content could affect the acoustic-phonetic values obtained. This analysis was possible because each talker pair completed three picture tasks per condition. To check for within-talker consistency, for each of the measures apart from the vowel ranges (which were calculated across three pictures to maximize the number of vowels measured), a repeated-measures ANOVA was carried out with picture (1st, 2nd, 3rd) and condition (NB, VOC, BABBLE) as within-subject factors, and gender as across-subject factor, for the 20 talkers recorded in all three conditions. The gender factor is not of particular interest per se (and is not reported) but is included to test for a picture by gender interaction; this would suggest that either men or women are less consistent in certain aspects of their speech production across pictures. The global measures of mean word duration, F0 median, F0 range, and ME1-3kHz were all found to be stable as shown by a lack of main picture effect or picture by gender interaction. It therefore seems that these gross acoustic-phonetic measures are stable within-condition even though the lexical content varied across each picture task. It is therefore likely that any differences in these acoustic-phonetic measures across condition are due to the condition itself rather than to the inherent variability that comes from the use of unscripted conversational speech as materials.

### C. Does the extent of acoustic-phonetic enhancements vary across the diapix "communication barrier" and read, clear tasks?

As many studies of clear speech have based their analyses on corpora involving read sentences with specific instructions given to speak clearly, it was of interest to see how the acoustic-phonetic characteristics of read, clear speech varied from those of speech produced in interaction between two talkers in adverse listening conditions. The VOC condition was chosen for this comparison as it was the communication barrier condition carried out by all 40 talkers. The two types of conversational speech (NB and read, conversational) were also included in the analysis to evaluate whether read, conversational speech was acoustically clearer than spontaneous, conversational speech. The amount of speech used in the comparison between conditions was of a similar order as, on average, the NH talkers produced 613 words (s.d. 191) in the NB condition and 759 words (s.d. 262) in the VOC condition. The sentence lists included 991 words.

For each of the acoustic-phonetic measures examined, repeated-measures ANOVAs were run with task type (diapix, read) and speech style (conversational, clear) as within-subject factors, and gender as between-subject factor, on the data obtained per talker, averaged across the three pictures per condition (see Table II). For median F0, the main effects of task type [F(1,38) = 24.1; $P < 0.001$] and speaking style [F(1,38) = 39.6; $P < 0.001$] were significant, with no significant interactions: F0 median (expressed in semitones re 1 Hz) was higher in read speech (87.3 st) than in the diapix (86.4 st) speech and was higher in the clear (87.4 st) than in the conversational (86.3 st) speech. The between-subject effect of gender was significant, as expected. For F0 range, the results were more complex. There was a significant task type by speaking style interaction [F(1,38) = 6.4; $P < 0.05$]

TABLE II. Median F0 (in semitones re 1 Hz), F0 range (interquartile range in semitones re 1 Hz), mean energy in the mid-frequency region of the long-term average spectrum (in dB), mean word duration (in ms) and vowel F1 and F2 range (in ERB) for male (N = 20) and female (N = 20) talkers in the diapix NB and VOC conditions, and the two conditions involving read sentences (read, conversational and read, clear).

| | Diapix, NB | | Diapix, VOC | | Read, Conv. | | Read, Clear | |
|---|---|---|---|---|---|---|---|---|
| | Female | Male | Female | Male | Female | Male | Female | Male |
| F0 median | 91.5 | 80.4 | 92.2 | 81.6 | 92.3 | 81.0 | 93.1 | 82.7 |
| (semitones) | (1.1) | (1.7) | (1.2) | (2.0) | (1.4) | (1.7) | (1.4) | (2.8) |
| F0 range | 3.3 | 2.8 | 3.2 | 3.1 | 3.1 | 3.4 | 3.6 | 4.1 |
| (semitones) | (0.8) | (0.6) | (0.7) | (0.6) | (0.8) | (0.9) | (0.9) | (1.0) |
| Mean energy | 25.6 | 23.6 | 27.5 | 25.4 | 22.9 | 23.4 | 23.4 | 25.7 |
| 1–3 kHz (dB) | (2.8) | (2.0) | (2.8) | (2.0) | (4.1) | (2.7) | (6.1) | (2.1) |
| Mean word | 265.9 | 250.0 | 345.1 | 310.2 | 252.4 | 246.6 | 412.8 | 427.9 |
| duration (ms) | (28.6) | (29.1) | (45.0) | (46.6) | (20.7) | (27.1) | (60.3) | (103.5) |
| F1 range | 3.5 | 3.8 | 4.7 | 4.4 | 3.8 | 4.0 | 5.0 | 5.3 |
| (ERB) | (1.0) | (0.7) | (1.1) | (0.7) | (1.1) | (0.6) | (1.1) | (1.0) |
| F2 range | 5.0 | 3.5 | 6.1 | 5.3 | 6.0 | 5.7 | 7.9 | 8.1 |
| (ERB) | (0.8) | (0.6) | (0.9) | (0.8) | (0.8) | (0.7) | (0.8) | (1.0) |

with a greater increase in F0 range between the conversational and clear styles in the read speech than in the diapix speech. There was a significant task type by gender [F(1,38) = 7.9; $P < 0.01$]: Men showed a greater increase in F0 range in clear speech produced in the read task than speech produced in the diapix task than women. Further, there was a style by gender interaction [F(1,38) = 4.2; $P < 0.05$], with a greater increase in F0 range for clear speech in men than in women. In terms of fundamental frequency measures, therefore, it seems that talkers speak with higher pitch when reading than in conversation, and that, in both types of speech, they produce speech with higher pitch in both the "imagined" and "real" challenging conditions. The increase in pitch range was more marked in the read, clear speech than in the diapix speech produced in the VOC condition especially for male talkers.

For MWD, the data analysis was carried out on the log-transformed data due to unequal variances in the raw data. There was a significant task type by speech style interaction [F(1,38) = 43.6; $P < 0.001$]: MWD did not differ between diapix and read speech in the NB/conversational condition (257 vs 249 ms, respectively) but was longer in the read, clear speech (420 ms) than in the VOC condition (328 ms). Talkers therefore slowed down their speech to a greater extent, or more consistently, when reading clearly than when clarifying their speech in interaction with another talker. The effect of talker gender was not significant.

For the mean energy 1–3 kHz (ME1-3kHz) measure, there was a significant task type by gender interaction [F(1,38) = 12.5; $P < 0.005$]: female talkers had more mid-frequency energy in their speech for the diapix (26.5 dB) than for the read task (23.2 dB) while the male talkers did not vary across task types (24.5 dB in both tasks). There was a main effect of speech style [F(1,38) = 17.4; $P < 0.001$] with higher ME1-3kHz in the clear speech (25.5 dB) than the conversational speech (23.9 dB). No other effects or interactions were significant. As there were a number of outliers, statistics were rerun with these removed. The effects present in the main data set remained, and the effect of gender was just significant [F(1,33) = 4.3; $P < 0.05$], with, on average, higher energy in the mid-frequency region of the female talkers' speech. Increases in mid-frequency energy therefore occurred for both the read, clear speech and the diapix speech produced in the VOC condition.

Vowel space was examined in terms of the F1 and F2 range expressed in Equivalent Rectangular Bandwidth (ERB) units. In terms of F1 range, there was a significant style by type interaction [F(1,37) = 6,14; $P < 0.05$]: There was a greater difference in F1 range between the conversational and clear conditions for the read speech (3.9 vs 5.2 ERB, respectively) than for the diapix speech (3.6 vs 4.5 ERB). The effect of gender was not significant although there was a three-way interaction among task type, style, and gender [F(1,37) = 5.9; $P < 0.05$]. For F2 range, the situation is more complex as there were significant interactions of task type with gender [F(1,37) = 28.9; $P < 0.001$]: men and women have a similar F2 range for the read speech (6.9 ERB), but women have a larger F2 range for the diapix speech (5.6 ERB) than the men (4.4 ERB). There was also a significant speaking style by gender interaction [F(1,37) = 11.0; $P < 0.005$]: Both men and women produced similar F2 ranges in the clear conditions (6.7 and 7.0 ERB, respectively), but men produced a more reduced range in the conversational condition (4.6 ERB) than did women (5.5 ERB). Finally, there was a significant task type by speaking style interaction [F(1,37) = 10.7; $P < 0.005$] with a bigger difference in F2 range between the read, conversational and read, clear speech conditions (5.9 vs 8.0 ERB, respectively) than between diapix NB and VOC conditions (4.3 vs 5.7 ERB, respectively). The picture regarding vowel production is therefore as follows: There was a tendency for the effect of speaking style on vowel range to be greater in read speech than in diapix speech, and women tended to use a more expanded vowel range in conversational speech than men.

In summary, there was evidence that read speech was produced with higher median F0 than in both the NB and VOC diapix conditions. Read speech also showed a greater change in F0 range between the conversational and clear conditions than did the diapix speech; this effect was more prevalent in men than in women's speech. The decrease in speaking rates across the conversational and clear conditions was also greater in read than in diapix speech. It is only for the measure of mid-frequency energy that there was evidence of greater energy in the diapix than read condition (and this, for women only). As the gender effects were found for measures that are most likely to be affected by physiological differences between men and women, despite our attempt to normalize for these differences by using semitone and ERB scales, we conclude that they are unlikely to signal significant gender-related differences in the strategies used in the production of clear speaking styles.

### D. Do spontaneous speech modifications vary according to the communication barrier?

The next set of analyses investigated whether talkers were able to adjust the acoustic-phonetic characteristics of their speech according to the type of communication barrier that their interlocutor was experiencing, even if they themselves were hearing normally (see Table III). Is it the case that talkers speak differently to a talker with a cochlear implant than they do to a talker who is hearing them by telephone in a noisy background for example?

In our experimental design, 20 talkers carried out the diapix tasks as the NH talker in the VOC and BABBLE conditions. Bearing in mind the likely individual talker differences in the overall acoustic-phonetic characteristics of their speech, a new metric was developed to look at the impact of the communication barrier on the speech produced: For each acoustic-phonetic measure and for each talker, the data for the communication barrier conditions were expressed in terms of the percent change relative to the speech produced by the same talker in the NB condition (i.e., relative to their conversational speech) (see Figs. 3 and 4). To investigate whether the acoustic-phonetic dimensions varied across the two communication barrier conditions, the BABBLE and VOC percent-change data were analyzed using a repeated-measures ANOVA with within-subject factors of condition

TABLE III. Median F0 (in semitones), F0 range (interquartile range in semitones), mean energy in the mid-frequency region of the long-term average spectrum (in dB), mean word duration (in ms), and vowel F1 and F2 range (in ERB) for the speech produced by the same group of male (N = 11) and female (N = 9) normal-hearing (NH) talkers in the NB, VOC, and BABBLE diapix conditions. Standard deviations are given in parentheses.

| | NB | | VOC | | BABBLE | |
|---|---|---|---|---|---|---|
| | Female (N = 9) | Male (N = 11) | Female (N = 9) | Male (N = 11) | Female (N = 9) | Male (N = 11) |
| F0 median | 91.4 | 81.1 | 91.8 | 82.3 | 93.1 | 85.3 |
| (semitones re 1Hz) | (1.3) | (1.5) | (1.3) | (2.3) | (0.7) | (2.1) |
| F0 range | 3.0 | 2.9 | 3.0 | 3.1 | 3.2 | 3.7 |
| (semitones re 1 Hz) | (0.7) | (0.7) | (0.9) | (0.7) | (0.8) | (0.7) |
| Mean energy | 25.2 | 23.7 | 27.3 | 25.5 | 29.7 | 27.9 |
| 1–3 kHz (dB) | (3.6) | (2.4) | (3.4) | (2.2) | (2.2) | (1.7) |
| Mean word | 276.7 | 251.8 | 359.5 | 311.1 | 350 | 322 |
| duration (ms) | (30.1) | (35.2) | (52.8) | (55.1) | (30.3) | (54.0) |
| F1 range | 3.7 | 3.6 | 4.9 | 4.4 | 5.3 | 4.7 |
| (ERB) | (1.0) | (0.5) | (0.9) | (0.7) | (1.1) | (0.6) |
| F2 range | 5.1 | 3.6 | 6.4 | 5.5 | 6.1 | 5.6 |
| (ERB) | (0.6) | (0.7) | (1.0) | (0.9) | (0.6) | (1.1) |

(BABBLE, VOC) and measure (F0 median and range, ME1-3kHz, MWD, F1 range, F2 range).

There was a significant condition by measure interaction [F(5,95) = 9.2; $P < 0.0001$], suggesting that the BABBLE and VOC conditions varied only for some measures. Paired-samples t-tests (see Table IV) carried out on the data for each acoustic-phonetic measure showed that talkers made greater changes in median F0 (20.1% for BABBLE vs 5.7% for VOC), F0 range (42.8% for BABBLE vs 10.1% for VOC), and ME1-3kHz (18.5% for BABBLE vs 8.5% for VOC) in the BABBLE condition than in the VOC conditions. Similar clarifications in the BABBLE and VOC conditions were made for MWD and vowel F2 range. For both measures, the degree of change was quite substantial, as MWD increased by an average of 27.1% in the VOC and 27.9% in the BABBLE conditions, and F2 range increased by an average of 41.9% in the VOC and 42.6% in the BABBLE condition. F1 range increased more in the BABBLE (38.7%) than VOC condition (28.4%).

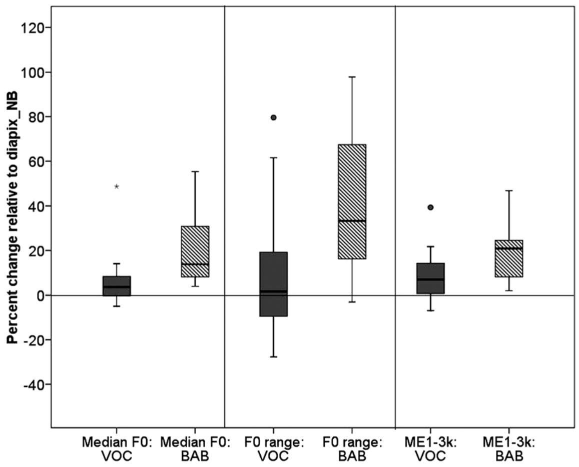

FIG. 3. Box-plots showing percent change in median F0, F0 range, and mid-frequency energy in the 1–3 kHz region (ME 1-3k) for the VOC (VOC) and BABBLE (BAB) conditions relative to the same talker's speech in the NB ("no barrier") condition.

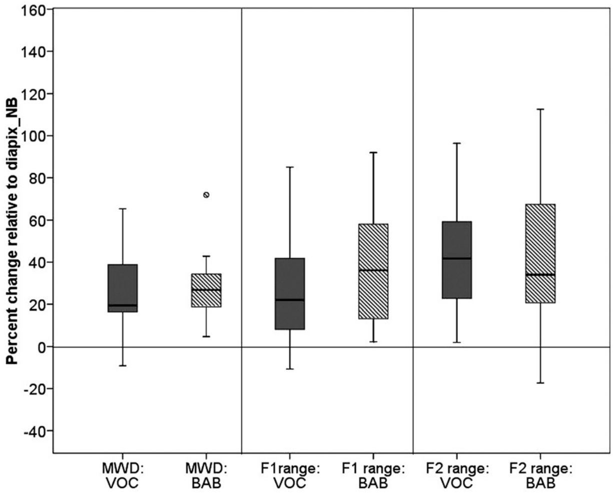

FIG. 4. Box-plots showing percent change in mean word duration (MWD), vowel F1, and vowel F2 range for the VOC (VOC) and BABBLE (BAB) conditions relative to the same talker's speech in the diapix NB ("no barrier") condition.

To check whether the cross-condition difference found for the F0 median and range, mid-frequency energy, and F1 range measures were due to differences in the extent of change relative to conversational speech or to these measures being enhanced in the BABBLE but not the VOC condition, repeated-measures ANOVAs were carried out on the raw data, with a within-subject factor of condition (NB, VOC, BABBLE) and between-subject factor of gender, for these four measures. These analyses confirmed that median F0 [F(2, 36) = 33.23; $P < 0.0001$] and F0 range [F(2, 36) = 8.66; $P < 0.001$] did not change significantly between the NB and VOC conditions. The ME1-3kHz measure [F(2, 36) = 45.39; $P < 0.0001$] and vowel F1 range [F(2, 36) = 28.22; $P < 0.0001$] were enhanced in the VOC condition relative to the NB condition, so the condition effect for these two measures when comparing the VOC and BABBLE conditions was one of extent of change.

The scatterplot in Fig. 5, showing the relation between percent change in F0 range (relative to the NB condition) for the VOC and BABBLE conditions for the 20 talkers, gives a picture of how the difference across conditions interacts with

TABLE IV. Paired-samples t-tests on the measures of percent change (relative to the NB condition) in F0 range, median F0, mean word duration, mid frequency energy (ME 1-3 kHz), F1 and F2 range in the VOC and BABBLE conditions for the group of 20 NH talkers.

| | t | df | Sig. (2-tailed) |
|---|---|---|---|
| Median F0 | −5.153 | 19 | 0.000 |
| F0 range | −5.097 | 19 | 0.000 |
| Mean word duration | −0.198 | 19 | 0.845 |
| ME1-3kHz | −6.911 | 19 | 0.000 |
| F1 range | 2.334 | 19 | 0.031 |
| F2 range | 0.134 | 19 | 0.895 |

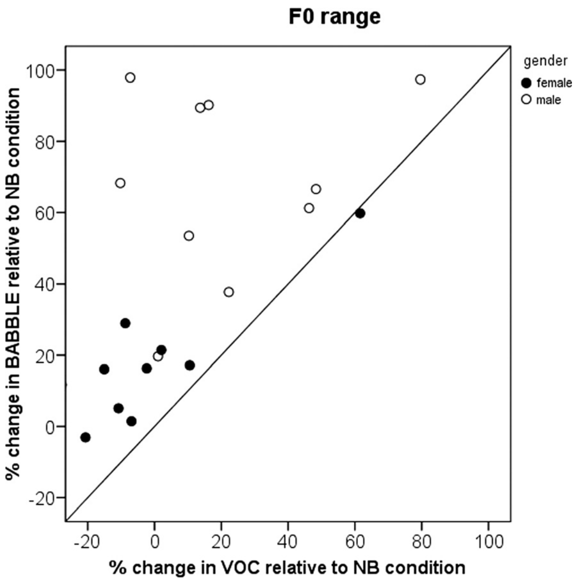

FIG. 5. Scatterplot showing the percent change (relative to the NB condition) for F0 range in the VOC and BABBLE conditions for individual talkers. For all but one talker, there is greater change in F0 range for the BABBLE than for the VOC condition.

individual talker strategies. There is certainly evidence of individual strategies in the acoustic-phonetic measures used by talkers to enhance their speech: Although on average F0 range did not change between the NB and VOC conditions, four talkers do seem to change their range substantially in the VOC condition as well as in the BABBLE condition. It is notable though that all but one talker show a greater change in F0 range in the BABBLE than in the VOC condition, so there is a general strategy to make greater changes to pitch range in the BABBLE condition.

In summary, it does appear that the strategies used by talkers who are themselves hearing normally but are interacting with interlocutors experiencing adverse listening conditions do vary according to the adverse listening condition. Even though the BABBLE and VOC condition were similar in difficulty, as measured by transaction time, talkers made greater changes in mean energy and F1 range in the BABBLE condition than in the VOC condition. Moreover, talkers did not change their median F0 and F0 range in the VOC condition relative to their conversational speech.

## IV. DISCUSSION

This study investigated the acoustic-phonetic characteristics of speaking styles used by a talker when interacting with an interlocutor experiencing an adverse listening environment. It investigated whether speech modifications made in such communicative conditions differ from clear speaking styles elicited through instruction in read speech materials for the same set of talkers. Further, we investigated whether speaking styles elicited via different communication barriers varied in acoustic-phonetic characteristics even though the talker whose speech is being analyzed was not experiencing the communication barrier directly.

The comparison between the read, clear condition and diapix VOC condition over 40 talkers showed that clear speech that is elicited via instruction in read sentences shows more extreme changes in at least certain acoustic-phonetic characteristics than speech produced to counteract intelligibility-challenging conditions. The extent of changes seen in clear speaking styles may also be influenced by the choice of talkers. For example, Picheny *et al.* (1985) selected talkers who had experience of public speaking (e.g., amateur dramatics), while Krause and Braida (2004) specifically trained their talkers over a period of an hour so that they could produce clear speech at a normal rate. These talkers are therefore likely to be at the "clear" end of a continuum of inherent talker clarity that is found in a normal population (Hazan and Markham, 2004; Bradlow *et al.*, 1996).

One probable reason for the greater enhancement of acoustic-phonetic measures found in clear, read speech is that the instruction to speak clearly induces the talker to produce a relatively consistent degree of clarification while reading a set of sentences in a laboratory setting. Within the process of communication between two talkers, however, as suggested by Lindblom's H&H model (1990), there is likely to be a constant tension between the need to clarify speech to ensure efficient communication and that of minimizing talker effort. Indeed, in the diapix recordings, it was apparent that when communication was occurring efficiently (i.e., without misunderstanding) in the communication barrier conditions, the degree of clarification reduced, although it increased again when a miscommunication occurred. The difference between the two speaking styles is therefore likely to be quantitative rather than qualitative and to be related to the proportion of speech that is enhanced. As random samples from the VOC and BABBLE conditions were judged to be clearer than samples from the NB condition, it is likely that the proportion of speech that is enhanced is substantial. In future, methods of analyses for spontaneous discourse that can more directly relate miscomprehensions and degree of speech enhancement would be more informative than the current approach of analyzing speech collected over the whole task.

It should be noted that although more natural than speech collected from read sentences, the type of speech that is elicited through the use of the diapix task is still far from natural communication. It is a referential task rather than spontaneous speech; the content of the speech is constrained in topic and syntactic complexity, and the laboratory setting introduces a certain level of formality to the exchanges. However, it constitutes a bridge between laboratory read speech and natural spontaneous speech by introducing communicative intent in the speech produced, and enabling researchers to collect speech that is relatively consistent across talkers in terms of the style and lexical content of the speech produced. Referential tasks that share similarities with the diapix task, in that they involve a transfer of specific information between two talkers, are frequently used in studies of communication with clinical populations (e.g., Leinonen *et al.*, 1997; Lloyd *et al.*, 2005).

The main aim of the study was to establish whether the acoustic-characteristics of the speech produced varied across different communication barrier conditions, and therefore whether there is evidence that talkers can adapt their speech to the needs of their interlocutor. Further, we can consider whether the changes being made are likely to promote greater intelligibility given the specific communication barrier that the interlocutor was experiencing.

In the BABBLE condition, F0 median, F0 range, mean energy, and F1 range were the characteristics that showed greater change in the BABBLE condition than the VOC condition, even though the two conditions were of similar difficulty as measured by transaction time. It does seem therefore that the changes made by NH talkers to clarify their speech for their interlocutors' hearing in noise share many characteristics with Lombard speech even though our NH talkers were not directly experiencing the noise heard during the task. In the VOC condition, we predicted that there would be little change in pitch characteristics and loudness as these would not be likely to aid intelligibility but that speaking rate and vowel space characteristics would change to a similar extent as in the BABBLE condition. Our findings generally confirmed these expectations: Talkers made no changes in their pitch characteristics relative to their conversational speech in the NB condition, and increased their mid-frequency intensity less than in the BABBLE condition but made similar changes to their speaking rate and vowel F2 range.

Overall, therefore, there is evidence that the trends seen in the speech modifications made in the two communication barrier conditions are well-matched to the needs of the AL interlocutor. This evidence gives further weight to the studies that have suggested that clear speech varies in characteristics according to the needs of the interlocutor (e.g., Uther *et al.*, 2007; Burnham *et al.*, 2002). Despite this trend, the wide variance typically seen in the various measures do indicate a significant degree of variability in the strategies used by individual talkers as has also been suggested in many previous studies of clear speech using read materials (Gagné *et al.*, 1994; Ferguson, 2004; Ferguson and Kewley-Port, 2007; Smiljanić and Bradlow, 2005) as well as in studies of Lombard speech (e.g., Summers *et al.*, 1988; Junqua, 1996). This is currently the subject of further analysis.

If it is the case that talkers who are themselves hearing normally can attune their clear speech characteristics to the needs of their AL interlocutors, the next question to ask is how talkers manage to make the most useful adjustments without simultaneously experiencing the communication barrier. Four possibilities may be envisaged. NH talkers may be getting cues as to the enhancements that would be most useful from the AL interlocutor and may adjust their speech through a process of phonetic convergence or accommodation, similar to the type of phonetic accommodation that is seen between talkers of different accents (e.g., Pardo, 2006; Delvaux and Soquet, 2007). A second possibility is that NH talkers are using their inherent knowledge of the communication barrier to guide the adjustments that will be most useful. This may be envisaged for the BABBLE condition and, to a certain extent, for the VOC condition as individuals had had a short period of training with vocoded speech prior to the diapix recordings although no explicit explanations were given as to the impact of the vocoder on speech. A third explanation is that the talker may use overt feedback from the AL talker to evaluate which adjustments are most helpful in terms of improving communication. A few examples of such overt feedback were found in our recordings (e.g., in the vocoder condition: "do not shout, it does not help"), but these are too few to explain how the adjustments were made. The final and most likely explanation is that the NH talkers were engaged in an ongoing process of trial and error in the specific speech clarifications being made, using covert feedback from the AL talker (i.e., which adjustments led to immediate understanding in the two-way exchange of information). This adjustment appears to be occurring very rapidly; indeed, the acoustic-phonetic measures made were stable across three 6-8 min long picture tasks within a given condition, suggesting that whatever condition-specific clear speech attunements were being made by the talker were primarily occurring within the span of the first of the three picture tasks. To verify this hypothesis, more micro-analyses are required to relate the acoustic-phonetic clarifications being made to a measure reflecting how successful the communication is between the two talkers (i.e., as marked by request for repetition, miscomprehension, etc.).

In conclusion, investigating the impact of communication barriers on speech in interaction between two talkers throws light on how talkers use the control that they have over their speech production to ensure effective and efficient communication. As expected, speech produced as a result of communicative need is more finely modulated than clear speech produced via elicitation and appears to be well-matched to the needs of the listener experiencing the adverse condition. It should be noted though that there are likely to be individual differences in the strategies used by talkers to clarify their speech in these different conditions and in the degree of success that individual talkers have in achieving effective communication. Further work is ongoing to explore the dynamic aspects of clear speech in interaction and its relation to discourse-related aspects of the communication and to explore how individual differences in the strategies used by talkers in clarifying their speech in different communicative conditions are related to communication effectiveness.

## ACKNOWLEDGMENTS

We wish to thank Ann Bradlow, external collaborator on this research project, and her colleagues from Northwestern University Linguistics Department for ongoing fruitful discussions about the diapix work and for granting us access to speech analysis software and other facilities at Northwestern University. We also wish to thank Martin Cooke for his valuable advice on the noise condition and for providing us with the babble noise used in our work, Andrew Faulkner for his help in implementing the real-time vocoder for the diapix sessions, Stephen Nevard for technical support, and Tom Neill, who drew the diapix pictures. This work was funded by the UK Economics and Social Research Council (RES-062-23-0681).

[1] The other 20 participants completed another (L2) condition in which they carried out the task with a nonnative talker. This condition is not reported here.

[2] The use of "friend" pairs for the NB and VOC condition but of "stranger" pairs for the BABBLE condition introduces a difference in familiarity across conditions. This was necessary as "strangers" were needed for the

L2 condition, not reported here, due to the extreme difficulty of finding L2 "friends" for each of the 40 participants that were matched in L2 proficiency; stranger confederates were also used for the BABBLE condition to have a balanced design between these two conditions. It was decided not to use stranger pairs for the NB and VOC conditions to avoid the potential effects of an increase in familiarity over the set of nine picture tasks carried out together. The fact that the VOC and BABBLE conditions did not vary in transaction time suggests that level of familiarity did not affect communication difficulty at least. Also, the difference between conditions involves greater F0 range and higher median F0 when speaking to a stranger (in the BABBLE condition); this counters the expectation that familiarity would lead to more animated conversation when talking to a friend.

[3] The complete LUCID corpus is accessible as part of the Online Speech/Corpora Archive and Analysis Resource (OSCAAR), based at the Northwestern University Department of Linguistics (http://oscaar.ling.northwestern.edu/, last accessed 31 March 2011), on request of password.

[4] For F1 range, the measures were typically based on the difference between F1 in /u:/ and F1 in /ae/. As an example of the amount of data on which the median range values were based, in the NB condition, for male talkers, the mean number of vowels measured to obtain the lowest median F1 for each talker was 89 (s.d. 43) and for the highest median F1, 66 (s.d. 24). For F2 range, the measures were typically based on the difference between the F2 in /o:/ and the F2 in /i:/. For the same subgroup of 20 male talkers, the mean number of vowels measured per talker for the lowest F2 median was 56 (s.d. 23) and for the highest F2 median 97 (s.d. 31).

[5] For the VOC condition, the analyses are based on the data obtained for 38 rather than 40 talkers as the first pair of talkers had been asked to write down rather than circle the differences found, thus distorting the transaction time measure.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (**1991**). "The HCRC Map Task Corpus," Lang. Speech **34**, 351–366.

Baker, R., and Hazan, V. (**2010**). "LUCID: a corpus of spontaneous and read clear speech in British English," Proceedings of DiSS-LPSS Joint Workshop 2010: 5th Workshop on Disfluency in Spontaneous Speech and 2nd International Symposium on Linguistic Patterns in Spontaneous Speech, University of Tokyo, Japan, September 25–26 September 2010, pp. 1–4 .

Bent, T., Buchwald, A., and Pisoni, D. B. (**2009**). "Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech," J. Acoust. Soc. Am. **126**, 2660–2669.

Boersma, P., and Weenink, D. (**2001**). "Praat, a system for doing phonetics by computer," Glot Int. **5**, 341–345.

Boersma, P., and Weenink, D. (**2010**). "Praat, a system for doing phonetics by computer," Version 5.1.32. www.praat.org (Last viewed 10/6/2010).

Bradlow, A. R. (**2002**). "Confluent talker- and listener-related forces in clear speech production," in *Laboratory Phonology*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin), Vol. 7, pp. 241–273.

Bradlow, A. R., and Bent, T. (**2002**). "The clear speech effect for non-native listeners," J. Acoust. Soc. Am. **112**, 272–284.

Bradlow, A. R., Kraus, N., and Hayes, E. (**2003**). "Speaking clearly for children with learning disabilities: Sentence perception in noise," J. Speech Lang. Hear. Res. **46**, 80–97.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech. I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.

Burnham, D., Kitamura, C., and Vollmer-Conna, U. (**2002**). "What's new pussycat? On talking to babies and animals," Science **296**, 1435.

Cooke, M., and Lu, Y. (**2010**). "Spectral and temporal changes to speech produced in the presence of energetic and informational maskers," J. Acoust. Soc. Am. **128**, 2059–2069.

Davis, B. L., and Lindblom, B. (**2001**). "Phonetic variability in baby talk and development of vowel categories," in *Emerging Cognitive Abilities in Early Infancy*, edited by F. Lacerda, C. von Hofsten, and M. Heimann (Lawrence Erlbaum Associates, Mahwah, NJ), pp. 135–171.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A. G., Taylor, K.J., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psych. Gen. **134**, 222–241.

Delvaux, V., and Soquet, A. (**2007**). "The influence of ambient speech on adult speech productions through unintentional imitation," Phonetica **64**, 145–173.

Ferguson, S. H. (**2004**). "Talker differences in clear and conversational speech: Vowel intelligibility for normal-hearing listeners," J. Acoust. Soc. Am. **116**, 2365–2373.

Ferguson, S. H., and Kewley-Port, D. (**2002**). "Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **112**, 259–271.

Ferguson, S. H., and Kewley-Port, D. (**2007**). "Talker differences in clear and conversational speech: acoustic characteristics of vowels," J. Speech Lang. Hear. Res. **50**, 1241–55.

Fernald, A. (**2000**). "Speech to infants as hyperspeech: Knowledge-driven processes in early word recognition," Phonetica **57**, 242–254.

Fernald, A., and Kuhl, P. (**1987**). "Acoustic determinants of infant preference for motherese speech," Infant Behav. Dev. **10**, 279–293.

Gagné, J.-P., Masterson, V., Munhall, K. G., Bilida, N., and Querengesser, C. (**1994**). "Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech," J. Acad. Rehab. Aud. **27**, 135–158.

Garnier, M., Henrich, N., and Dubois, D. (**2010**). "Influence of sound immersion and communicative interaction on the Lombard effect," J Speech Lang. Hear. Res. **53**, 588–608.

Hazan, V., and Markham, D. (**2004**). "Acoustic-phonetic correlates of talker intelligibility for adults and children," J. Acoust. Soc. Am. **116**, 3108–3118.

Huckvale, M. (**2008**). Speech Filing System, Version 4.7. www.phon.ucl.ac.uk/resource/sfs/. (Last viewed 10/6/2010).

Ibertsson, T., Hansson, K., Maki-Torkko, E., Willstedt-Svensson, U., and Sahlen, B. (**2009**). "Deaf teenagers with cochlear implants in conversation with hearing peers," Int. J. Lang. Commun. Disord. **44**, 319–337.

Johnson, K., and Mullennix, J. W., eds. (**1997**). *Talker Variability in Speech Processing* (Academic, San Diego), 237 pp.

Junqua, J.-C. (**1996**). "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," Speech Commun. **20**, 13–22.

Kain, A., Amano-Kusumoto, A., and Hosom, J. P. (**2008**). "Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility," J. Acoust. Soc. Am. **124**, 2308–2319.

Krause, J. C., and Braida, L. D. (**2004**). "Acoustic properties of naturally produced clear speech at normal speaking rates," J. Acoust. Soc. Am. **115**, 362–378.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (**1997**). "Cross-language analysis of phonetic units in language addressed to infants," Science **277**, 684–686.

Lane, H., and Tranel, B. (**1971**). "The Lombard sign and the role of hearing in speech," J. Speech Hear. Res. **14**, 677–709.

Leinonen, E., and Letts, C. (**1997**). "Referential communication tasks: Performance by normal and pragmatically impaired children," Eur. J. Dis. Commun. **32**, 53–65.

Lloyd, J., Lieven, E., and Arnold, P. (**2005**). "The oral referential communication skills of hearing-impaired children," Deaf. Educ. Int. **7**, 22–42.

Lindblom, B. (**1990**). "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic, The Netherlands), pp. 403–439.

Liu, S., Del Rio, E., Bradlow, A. R., and Zeng, F.-G. (**2004**). "Clear speech perception in acoustic and electrical hearing," J. Acoust. Soc. Am **116**, 2374–2383.

Liu, S., and Zeng, F.-G. (**2006**). "Temporal properties in clear speech perception," J. Acoust. Soc. Am. **120**, 424–432.

Lu, Y., and Cooke, M. (**2008**). "Speech production modifications produced by competing talkers, babble and stationary noise," J. Acoust. Soc. Am. **124**, 3261–3275.

Moon, S.-J., and Lindblom, B. (**1994**). "Interaction between duration, context, and speaking style in English stressed vowels," J. Acoust. Soc. Am. **96**, 40–55.

Pardo, J. S. (**2006**). "On phonetic convergence during conversational interaction," J. Acoust. Soc. Am. **119**, 2382–2393.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1985**). "Speaking clearly for the hard of hearing. I. Intelligbility differences between clear and conversational speech," J. Speech Hear. Res. **28**, 96–103.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing. II. Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446.

Rosen, S., Faulkner, A., and Wilkinson, L. (**1999**). "Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants," J. Acoust. Soc. Am. **106**, 3629–3636.

Smiljanić, R., and Bradlow, A. R. (**2005**). "Production and perception of clear speech in Croatian and English," J. Acoust. Soc. Am. **118**, 1677–1688.

Smiljanić, R., and Bradlow, A. R. (**2009**). "Speaking and hearing clearly: Talker and listener factors in speaking style changes," Linguist. Lang. Compass **3**, 236–264.

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., and Stokes, M. A. (**1988**). "Effects of noise on speech production: acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**, 917–928.

Tye-Murray, N. (**1995**). "Effects of talker familiarity on communication breakdown in conversations with adult cochlear-implant users," Ear Hear. **16**, 459–469.

Uchanski, R. M. (**2005**). "Clear speech," in *Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell Publishers, Malden, MA), p. 207–235.

Uchanski, R. M., Choi, S., Braida, L. D., Reed, C. M., and Durlach, N. I. (**1996**). "Speaking clearly for the hard of hearing. IV. Further studies of the role of speaking rate," J. Speech Hear. Res. **39**, 494–509.

Uther, M., Knoll, M. A., and Burnham, D. (**2007**). "Do you speak E-NG-L-I-SH? Similarities and differences in speech to foreigners and infants," Speech Commun. **49**, 1–7.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (**2010**). "The wildcat corpus of native- and foreign-accented english: Communicative efficiency across conversational dyads with varying language alignment profiles," Lang. Speech **53**, 510–540.

Wassink, A., Wright, R., and Franklin, A. (**2006**). "Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers," J. Phon. **35**, 363–379.