Proc. of ASRU 2003, Dec 2003, U.S Virgin Islands

ACOUSTIC CORRELATES OF USER RESPONSE TO ERROR IN HUMAN-COMPUTER DIALOGUES

Abe Kazemzadeh, Sungbok Lee, and Shrikanth Narayanan
University of Southern California, Los Angeles, CA 90089, USA
{kazemzad, sungbokl}@usc.edu, {shri}@sipi.usc.edu

ABSTRACT

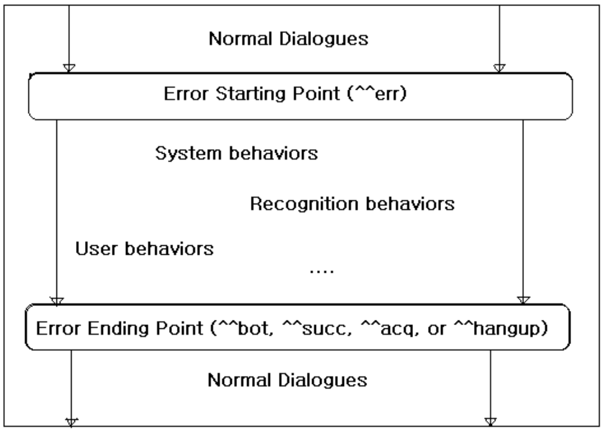
Using tagged data from the DARPA Communicator Project, we investigate acoustic features of user responses to system errors. We measure acoustic parameters such as energy, fundamental frequency, sub-band energy, ratios of voiced, unvoiced and silent regions of speech, fundamental frequency slope, spectral slope, and spectral center of gravity. We investigate different types of user responses to the errors, including frustration and various types of corrections. It is confirmed that the most prominent acoustic parameter for responses to the errors is fundamental frequency maximum and range, while other features are found to be salient for specific reaction types. More interestingly, acoustic characteristics of user responses to the errors are found to be different depending on whether the responses are the initial or continued responses to the errors. Similarly, normal user responses can differ acoustically depending on whether or not they were preceded by responses to error. We also present results on automatic classification of error response types using these features.

1. INTRODUCTION

Spoken dialogue systems promise a convenient, robust, and natural interface for human-computer interaction. Literature on human-computer interaction has detailed how the moods and emotional states of the users can influence their cooperation, attention span, and other faculties that affect their ability to work with the computer [1,7,9]. A major advantage of spoken dialogue systems is that the medium of spoken communication is a rich source of information regarding the user's interest, effort, and emotional state. Ideally, instead of banging on a keyboard, users can interact with such a spoken dialogue system with a variety of options ranging from changes in intonation to lexical choices to dialogue actions. Due to prevalence of errors in those systems and the negative effect of error on the human-computer interaction, however, we at SAIL lab (http://sail.usc.edu/) have undertaken several projects to study errors in human-computer dialogues.

In our previous first study [1] of the Communicator systems, we developed an account of user behavior under error conditions based on dialogue strategies of the user and system. By defining two states, error and not error, and marking them by manual tagging we succeeded in formulating general trends that can determine the successfulness of user and system attempts to correct errors in the dialogues. Figure 1 illustrates how we have utilized our tags to determine the location of error blocks, from where errors began to where they were either corrected, negotiated, acquiesced to, or abandoned.

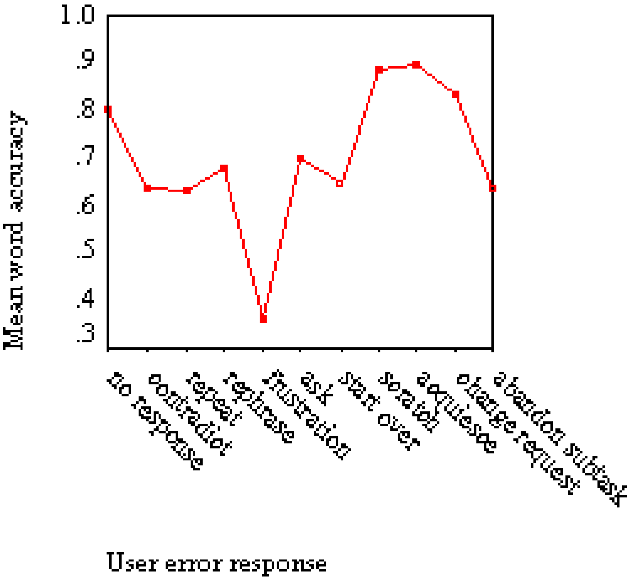
Figure1: the definition of an error block



Error block length and probability smoothed by an exponentially weighted word error rate (WER) measure provided measures of the effectiveness of different system and user strategies to get the dialogue back on track. By using tag information from manually tagged data and word error rate (WER) information based on manual transcriptions, we made use of higher level information that is not available in a real time dialogue. In this present paper, we examine acoustic features of users' responses to error in order to complement our previous approach with lower level acoustic information. These low-level features would provide a deeper insight into specific user state properties.

Previous studies [1,2] have shown that, in human-computer dialogues, users' reactions to errors have a varying degree of effectiveness in the correction of the errors. Moreover, it has been observed that the users switch to a hyper articulated speaking style when attempting to correct the system, resulting in further complications to the error state due to poorer recognition of these corrections [3]. In Figure 2, we have plotted the average recognition accuracy of different user behaviors, which will be fully described later.

Figure 2*:* The effect of error induced behavior on word accuracy. The word accuracy of most error responses is diminished compared with non-error responses ("no response").



The mean word accuracy in most of the cases in which the user is responding to perceived errors is significantly less than when there are not user reactions to any error (all results noted as *significant* are of the significant level of p < .05 in ANOVA, unless stated otherwise.). The exceptions are when the user attempts to correct an error by using the built in scratch or cancel turn command, when the user acquiesces to an error, and when they change the requested date or location of the arrival/departure. It is important to note that the responses to error in the table are not where the original ASR error of user utterance occurred, but rather the effects of user perception of system error after the error already occurred, therefore the recognition accuracy shown above, is not representative of what caused the error in the first place. However, this decrease of recognition accuracy does indeed cause more subsequent errors. This shows that there is a need to characterize the changes in the acoustics of speech when users respond to errors. In order to understand the dynamics of the user behavior in error conditions, we must clearly delineate the acoustic features of the users' responses to errors with reference to their function and position within the dialogue. This elucidation of the relation of speech acoustics with user dialogue behavior can be seen as a way of studying the prosody of human-computer spoken language interaction.

2. DATABASE

2.1 Corpus

We used audio data and tagged transcriptions from the Communicator Travel Planning Systems [4,5,6] June 2000 recording. Each dialogue consists of a number of exchanges between a computer travel agent and a human. Each transcribed exchange consists of a system utterance, a user utterance (manually transcribed from recordings), and what the ASR system heard and provided as input to the dialogue system.

Along with each user turn there was a corresponding audio file with the user's recorded utterance. These were recorded in NIST sound file format, encoded in either pcm or mu-law, at a sampling rate of 8000 Hz. These were aligned to the transcriptions, with manual corrections made to adjust occasional mismatches that resulted from empty sound files. In all, we analyzed 2586 utterances.

The data and the collection procedure are described in detail in [6]. In the Communicator dialogues, 85 experimental subjects interacted with 9 different "travel agent" systems. We worked with a subset of 141 dialogues and the average length of these dialogues was 18 exchanges.

2.2 Tagging

The aforementioned 141 dialogues were manually tagged by two annotators who showed 87% inter-annotator agreement. Our tagging scheme was devised with in order to highlight the ways in which errors are recognized and dealt with by the user [1]. The detailed tag set, exemplary user utterances and its usage conventions can be found in http://sail.usc.edu/dialog/model_tags.html and also in [7].

Briefly, the tag set included (1) SYSTEM tags: explicit confirmation, implicit confirmation, help, system repeat, reject, non sequitur, (2) USER tags: repeat, rephrase, contradict, frustration, change request, start over, scratch, ask, acquiesce, hang-up, (3) TASK tags: error, back on track, and task success. Since we are interested in characterizing the users' acoustic behaviors, we will focus on the USER tags.

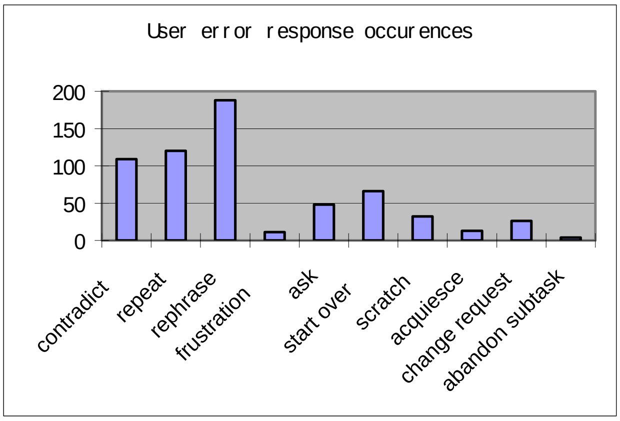
An important characteristic of our tagging is that it was tagged by reading the transcripts of the dialogues. Therefore, tags like frustration are not based on the perceived quality of the speech, but rather from information from the text transcriptions, such as swearing and the like. For example (snippets drawn from several different dialog sessions):

User said: i don't wanna do anything with you anymore ^^frust
User said: no you suck ^^frust

User said: /noise/ forget it canceling this call goodbye ^^frust
User said: quit ^^frust
User said: that is correct stop asking me yes it’s correct ^^frust
User said: i'll say it again yes yes yes ^^rephrase ^^frust

This approach of using data tagged by annotators reading transcriptions has been used in detecting corrections in [2], while approaches to detecting emotions have generally favored using human listeners [9] or actors [10] [11]. Figure 3 shows the histogram of the tagged user behaviors in our corpus. By taking our approach we were aiming to observe the low-level acoustic features of high-level dialogue strategies.

Figure 3: A histogram showing the number of error response types in our tagged corpus.



3. METHODOLOGY

3.1 Acoustic measurements

In order to investigate acoustic variations that are correlated with prosodic aspects of user utterances, we used 35 different parameters of speech related to energy, pitch, duration, voicing, and frequency spectrum of each utterance. They are listed in Table 1. A normalized range is also defined as

Normalized range = (max-min)/(max+min).

We used this parameter in order to minimize turn-to-turn variation of user utterances.

These measurements were obtained using Entropic Research Laboratory’s ESPS package. First, (after removing the DC component with *rem_dc*) we used the program *find_ep* to find the endpoints of the utterance. From the results of this program we determined the duration of the speech and the preceding and trailing silence. The energy, pitch and voicing measurements were obtained by using the ESPS program *get_f0*. The only the energy and pitch of the voiced speech were used (relying on the prob_voice feature). The percentages of voiced sounds, unvoiced sounds, and silence were measured by finding the ratio of the number of voiced, unvoiced, and silent (below 50 rms) records from *get_f0* out of the total number of records. The filters for the band-limited energy values were developed using the ESPS program *xfir_filt* and executed using *filter*. The frequency spectrum was obtained using ESPS’s *fft* program. The spectral center of mass was obtained by using the amplitude of each frequency of the fft as a weight and computing the weighted average. The f0 slope (as well as the fft slope) was calculated using a c++ regression class [8].

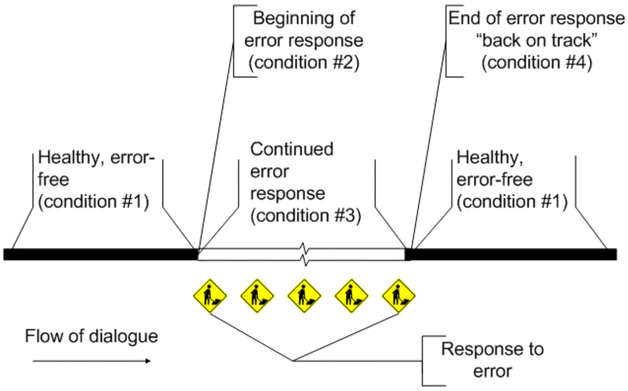
Table 1: Acoustic measurements performed in the study.

Energy Measurements:
RMS energy of voiced segments: minimum, maximum, range, and normalized range, Band-limited energy of voiced segments: minimum, maximum, range, and normalized range, Band 1: 0-1300 Hz Band 2: 1300-2600 Hz Band 3: 2600-4000 Hz
Pitch Measurements:
Fundamental frequency (f0): minimum, maximum, range, and normalized range. f0 slope (linear regression coefficients)
Duration Measurements:
Duration of the speech, duration of preceding silence, duration of end silence.
Voicing Measurements:
Percent voiced, percent unvoiced, percent silence.
Frequency Spectral Envelope Measurements:
Spectral slope: minimum, maximum, range, and normalized range. Spectral center of mass (moment): minimum, maximum, range, and normalized range.

3.2 Conditional error response environments

In addition to observing the acoustic features of response to error as it is labeled by our tags, we also observed the acoustic trends in four conditional environments which we defined by an utterance’s relation to the preceding utterance: 1) a non-error response is followed by another non-error response (i.e., a healthy, error-free dialogue turn), 2) an error response follows a non-error response (i.e., the beginning of a response to errors), 3) an error response that follows another error response (i.e., a continued error), and 4) a non-error response follows an error response (i.e., the end of an error response). We felt that observing these environments would be advantageous to see if continued response to error will show any differences with the initial response or if there is any marked difference between non-error responses depending on whether or not they are immediately preceded by error responses. Moreover, significant differences in these environments would aid in detecting regions of error.

Figure 4: Diagram illustrating the conditional environments that were analyzed separately, in addition to the error response types marked by tags.

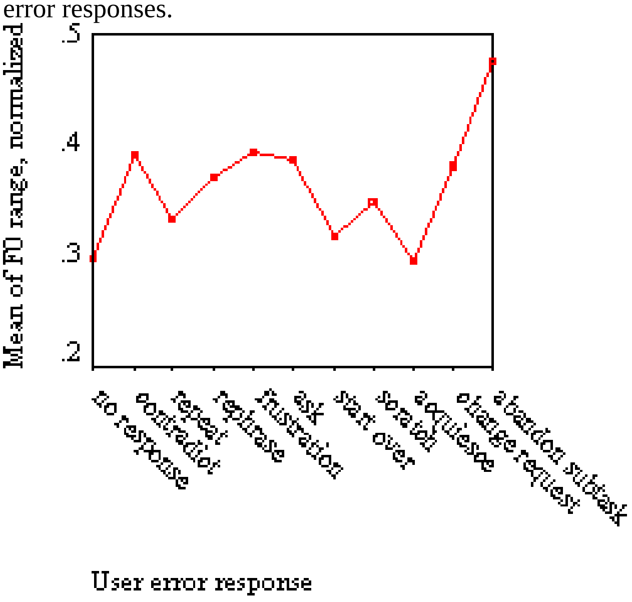

Beginning of error response (condition #2)
End of error response "back on track" (condition #4)
Healthy, error-free (condition #1)
Continued error response (condition #3)
Healthy, error-free (condition #1)
Flow of dialogue
Response to error

4. RESULTS

4.1 General features of response to error

As shown in Figure 5, the main acoustic feature shared by most user behavior in response to error is higher fundamental frequency maximum and range. This holds true significantly for contradict, rephrase, and frustration.

Figure 5: Normalized F0 range values for different user error responses.

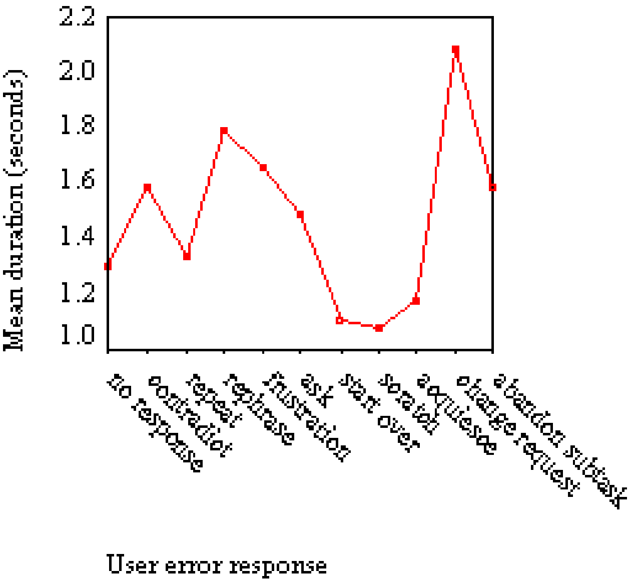


In addition to fundamental frequency information, the increased duration of utterances, as shown in Figure 6, proved to be a general feature of reactions to error, except in inherently short built-in utterances such as start over, scratch, and acquiescence. This trend was only statistically significant for contradict and rephrase, however.

4.2 Features of specific response types

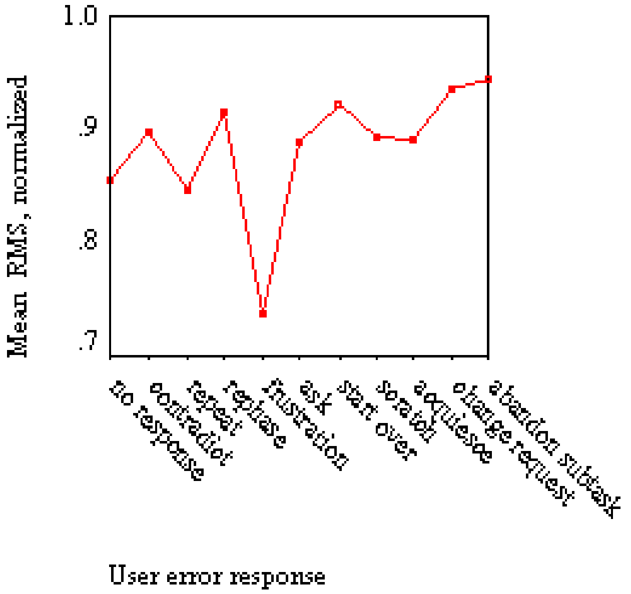
Among the different types of user responses to error there are differences in acoustic features. Normalized RMS energy range in particular, shown in Figure 7, shows much variation among different types of response to error.

Figure 6: Mean Utterance durations of error response types.



The mean normalized RMS energy range of frustration is significantly lower than untagged utterances, contradiction, rephrase, ask, start over, and scratch. Also, contradict and rephrase stand out as corrections types which have significantly more energy than a non-error response as well as when the user simply repeats the previous utterance.

Figure 7: Mean RMS energy of error response types.



4.3 Conditional error environments

One of the significant aspects of the conditional error environments is the difference between non-error responses that come immediately after an error response (i.e., "back on track", condition #4 in Figure 4) and those that do not (error-free dialogue, condition #1 in Figure 4). Utterances in error-free dialogue regions are significantly greater in energy than turns when the dialogue gets back on track after errors being successfully corrected in the previous turn. Also non-error responses in back on track positions have less percentage of voiced speech and a greater fft slope maximum than non-error responses in completely error-free regions. This could be seen as evidence for a period of relief after the user has spent energy on correcting an error.

There are also acoustic differences between initial error responses (condition #2 in Figure 4) and continued error responses (condition #3 in Figure 4). One significant difference is that the RMS maximum is higher for initial responses to error than for continued responses. Also the change in RMS range from the previous turn is higher for initial errors. However, the F0 maximum and range are significantly higher for continued errors than for initial errors. This could be interpreted as evidence for a strong, perhaps angry, initial response to error giving way to frustration or resignation.

The benefit of distinguishing these environments using properties of the user's speech signal is that it would enable precise location of dialogue states using a dialogue model that is based entirely on user responses.

4.3 Discriminant analysis

To see how effectively these measurements could be used in automatic classification, Fisher's linear discriminant analysis is performed in order to test our set of features. First we tested effectiveness of the acoustic features to predict response to error versus non-error responses. Using the full feature set, the 64.3% of the cases were correctly classified. Testing the classification of the four conditional environments the classification accuracy was 42.2%. When classifying the each individual tag, the discriminant analysis yielded 24.1% classification accuracy. Although these classification scores are low, they are well above the baseline of random choice. Adding other features in addition to the acoustic features improved the classification accuracy a little bit. Such features that we added were word accuracy (WA), speech rate, and dialogue length at time of utterance. These results are summarized in Figure 8.

classification groups	Feature sets			
	acoustic features	acoustic plus WA	acoustic plus sp. rate	acousticp lus dial. Length
Error response or not (binary)	64.3%	69.3%9	64.2%	65.9%
Conditional environments (four-way)	40.6%	46.5%	42.4%	43.9%
individual tags (11-way)	24.1%	31.1%	24.4%	27.0%

Figure 8: Results of discriminant analysis.

5. DISCUSSION

In this research we tested the hypothesis that, in dialogue systems, different categories of user response to error would show discernable acoustic differences. The current study indicates such a tendency and these results could improve error detection and thereby the robustness of dialogue management.

One interesting implication of this research is the correlation with emotion (i.e., the user's attitudinal state). Although human-computer dialogues are far from natural interpersonal communication, we can compare the results of our measurements with traditional descriptions of emotions. For example, "frustration," as we tagged and measured it, has more in common with traditional descriptions of resignation and sadness than anger [11]. This is shown most significantly in the lower normalized range of energy, lower energy range in the high frequency band, and the lower percentage of voiced speech. "Contradiction" bears most similarity with anger because of the increased F0 and RMS maximum and range. Also the increased percentage of voiced speech with the decreased percent silence gives evidence of increased speaking rate, which we confirmed by estimating the speech rate by dividing the number of words from the transcription by the duration (other studies have estimated speech rate using purely acoustic features by counting the number of voiced segments and dividing by time [3]). In addition to contradiction, rephrasing and asking both have higher speech rates.

Another implication of this research is that it may be helpful to use a measure of user effort to model user interaction with dialogue systems. Different user responses to error seem to imply varying levels of effort and initiative. For example, the "contradict" and "rephrase" user responses have relatively high mean f0 and RMS energy ranges and long durations when compared with repeat. However, though these measurements correlate with user effort, not all user effort is beneficial. In our previous research [1] and from word accuracy statistics (Figure 1), it can be seen that rephrasing information is helpful when users respond to errors to errors, but our previous research has shown contradiction to be a less successful user error response.

Modeling the relation of user emotion, effort and initiative with error response types is one direction of future research in this area. The other direction is combining the acoustic cues with lexical and other non verbal cues such as confidence measures in recognition level as well as in semantic level for improved assessment of the user's state under error conditions. Of course, online acquisition and process of acoustic and higher-level information in real-time for dialogue management purpose are not trivial tasks and it poses another technical challenge.

6. REFERENCES

[1] J. Shin, S. Narayanan, L. Gerber, A. Kazemzadeh, and D. Byrd, "Analysis of User Behavior Under Error Conditions in Spoken Dialogues", *Proc. of ICSLP*, Denver, 2002.

[2] M. Swerts, D. Litman, and J. Hirshberg, "Corrections in Spoken Dialogue Systems", *Proc. of ICLSP*, Beijing, 2000.

[3] J. Hirschberg, D. Litman, and M. Swerts, "Identifying User Corrections Automatically in Spoken Dialogue Systems", *NAACL*, Pittsburg, 2001.

[4] W. Ward and B. Pellom, "The CU communicator system", *IEEE ASRU*, Keystone, CO, 1999.

[5] E. Levin, S. Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. di Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker, "The AT&T-DARPA Communicator mixed initiative spoken dialog system", *Proc. of ICSLP*, Beijing, 2000.

[6] M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker, "DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection", *Proc. Eurospeech*, Aalborg, 2001.

[7] S. Narayanan, "Toward modeling user behavior in human-machine interactions: Effects of Errors and Emotions", *ISLE Workshop on Multimodal Dialog Tagging*, Edinburgh, UK, Dec 2002.

[8] D. Swaim II, "A simple c++ regression class", *C/C++ User's Journal*, http://www.cuj.com/, viewed 4/20/03.

[9] C. Lee and S. Narayanan, "Towards Detecting Emotion in Spoken Dialogs", *IEEE Transactions on Speech and Audio Processing*,

[10] J. Noad, S. Whiteside, and P. Green, "A macroscopic analysis of an emotional speech corpus", *Proc.of Eurospeech*, Rhodes, Greece, 1997.

[11] I. Murray and J. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", J. Acoust. Soc. Am., 93 (2), Feb. 1993.