Encoding and decoding Confidence information in speech

Xiaoming Jiang [1], *Marc D. Pell* [1]

[1] School of Communication Sciences and Disorders and Center for Research on Brain, Language and Music, McGill University, Canada

xmjiang1983@gmail.com, marc.pell@mcgill.ca

10.21437/SpeechProsody.2014-103

Abstract

This study aims to investigate the perceptual-acoustic correlates of vocal confidence. Statements with different communicative functions (e.g., stating facts, making judgments) were spoken in confident, close-to-confident, unconfident and neutral voices. Statements with preceding linguistic cues (e.g. I'm positive, Most likely, Maybe, etc.) or no linguistic cues were presented to sixty listeners in a perceptual study. The listeners were asked to judge whether statements conveyed some level of confidence, and if so, they were asked to evaluate the level of confidence of the speaker. The results demonstrated that the intended levels of confidence varied in a graded manner in the perceptual rating score; the more confident the statement intended to be, the higher the rating. In general, the neutral voice was judged to be more confident than the close-to-confident voice, but less than the confident voice. The presence of a linguistic cue tended to increase ratings of confident voices but decrease ratings of voices in the less confident voice conditions. To evaluate how specific prosodic cues are used to encode and decode confidence information, acoustic analyses were performed on the stimuli without the linguistic cue based on the mean perceptual rating of speaker confidence for each item. Results showed that statements rated as confident versus unconfident differed in the mean and the variance of fundamental frequency (f0) as well as speech rate, with confident statements exhibiting lower mean f0, smaller f0 variance, and faster speaking rate than unconfident statements. The perceived level of confidence was differentiated in the mean fundamental frequency in a parametric way, the lower the level of confidence, the higher the mean f0. Confident voices were also distinct from the other three conditions in terms of mean and range of amplitude (i.e., loudness). These findings shed light on how linguistic and paralinguistic cues reveal confidence-related information to listeners during speech.

1. Introduction

Humans have the ability to encode different types of emotive and social meanings in speech communication, which are often understood by listeners through an inferential process that weighs evidence from available linguistic and paralinguistic cues. Among the different emotive states that can be expressed are emotive devices that serve to foreground speaker-content (*evidentiality* devices). Within this category, *confidence* refers to cues that provide evidence of the reliability, correctness, or truth value of a speaker's statement; in social interactions, listeners make inferences about the confidence of other speakers to make appropriate decisions based on the perceived reliability of what is said, and they may also use this information to associate specific social or personality traits to the speaker (e.g. perceived confidence is usually associated with persuasiveness, [1]). It is suggested that evidentiality/confidence is communicated through the choice of specific linguistic structures (e.g. modal adverbs) and/or the speaker's prosody (e.g. changes in pitch / intonation contour and other acoustic parameters to make speech sound doubtful, certain, authoritative, submissive, etc.). However, little empirical work has been conducted on how different levels of confidence are realized by a speaker in terms of prosodic variation, nor is it well known how speakers use prosodic and linguistic cues in the decoding of speaker confidence. The goal of the present study was to supply preliminary data on perceptual and acoustic features of utterances that convey varying levels of confidence in English, as a first step for advancing knowledge of how evidentiality devices operate in speech communication.

2. Methods and results

2.1 Emotion elicitation study

2.1.1 Participants

Six native Canadian English speakers (Mean age in years = 22.8, three females) were recruited to produce statements expressing different levels of confidence in their native language. Speakers were selected for having lay experience in acting (e.g. in community theatre) or in public speaking (e.g. radio) and were compensated $35 after the recording session.

2.1.2 Materials

One hundred and fifty-one sentences were constructed by a native speaker of Canadian English (one of the authors, MDP). Sentences were of three types: 67 were statements describing a fact, 40 were descriptions of one's intention to initiate an action, and 44 were descriptions of one's judgment toward other people or things. Each sentence was intended to be produced in a neutral manner and to convey three levels of confidence about the subject matter: confident, close-to-confident, and unconfident. In order to facilitate production of sentences that were inflected to convey different levels of confidence in a way that was as natural as possible, lexical phrases consistent with each level of confidence were used as the beginning of the sentence during recording. For the confident level, speakers began with either "Definitely", "For sure", "I'm certain" or "I'm positive"; for the close-to-confident level, they produced "I think" "Most likely", "I'm pretty sure", "I'm almost certain"; and for the unconfident level, they began with "Maybe", "Perhaps", "It's possible", "There's a chance". Finally, 151 wh-questions were also created to facilitate naturalistic expressions of each sentence during the recording session by having speakers produce target sentences as part of a mini-dialogue (e.g. What will happen? – We will run out of gas). Care was taken that none of the questions led speakers to emphasize specific constituents in

the sentence, but rather, they could freely respond to the question with the appropriate level of confidence.

2.1.3 Elicitation and recording procedure

Each speaker was recorded separately in a sound-attenuated chamber. Sentences conveying neutral affect and each of the three levels of confidence were recorded in a separate block during the elicitation study. The neutral level always preceded the other levels. The order for recording specific confidence levels was randomized across speakers. Within each confidence condition, the three types of sentences (facts, intentions, judgments) were also recorded in a separate block. The order for recording specific type of sentences was randomized across confidence levels.

To facilitate expressions of confidence that were as naturalistic as possible, we created a dialogue setting in which the actor responded to questions posed by a female examiner, who was a native Canadian English speaker. For each confidence level, the examiner asked the question constructed for each target sentence and then the speaker produced the sentence with the appropriate level of confidence as if answering the speaker's question. The examiner provided clues to the speaker at the onset of each level that included descriptions of a scenario that would be likely to elicit the target level of confidence [2]. However, at no time did the examiner never model the vocal features that may be associated with specific impressions of confidence to participants. In the confident/close-to-confident/unconfident recording blocks, the speaker was instructed to produce the lexical phrase followed by the main sentence stem and to try to portray the level of confidence in their voice (prosody) throughout the sentence. They repeated each sentence twice. Recording blocks were separated by a short break to help the transition between modes of social expression. After producing each confidence level, speakers were asked to rate on a 7-point confidence scale (1=not at all confident, 7= very confident) how confident they were subjectively feeling when they produced sentences in each condition; the mean ratings for the six speakers were 6.5 in the confident voice condition, 4.7 in the close-to-confident condition, 1.8 in the unconfident condition, and 5.0 in the neutral condition.

All utterances were recorded onto digital media (Tascam Recorder) using a high-quality head-mounted microphone. The recordings were transferred to a computer and were saved as individual sound files in Praat. To reduce the number of exemplars included in the perceptual rating study, the two repetitions of each item produced by a given speaker were initially evaluated by two native English speakers to select the best (single) exemplar per item/speaker, based on a judgment of which item conveyed the intended target level of confidence, while excluding items that sounded unnatural (i.e., posed) and/or that had recording artifacts. This process yielded 3624 stimuli in total (6 speakers x 4 confidence levels x 151 items).

2.2 Perceptual study

2.2.1 Participants

A total of 60 listeners took part in the study (31 females and 29 males, with the mean age 25.2 years, mean education 16.6 years). All listeners were born and grew up in Canada and had English as their first language. None had lived outside of Canada for more than 1 year. All participants had normal hearing and no history of psychiatric or neurological disorders.

2.2.2 Materials and procedure

To test the perceptual recognition of confidence from utterances that contained lexical phrases or just prosodic cues, all recordings were further edited by removing the lexical phrase; this produced two versions of each statement with and without the lexical phrase ("with-cue" and "no-cue statement"). The resulting 6342 statements (6 speakers x 4 confidence levels x 151 items 2 cue types) were divided into six experimental lists (each with 1057 statements). The "with-cue statement" and the "no-cue statement" with the same sentence stem produced by the same speaker in the same prosody were never repeated in one list. Statements in each list were randomized for each listener and were separated in 20 short blocks. Each list was judged by 10 participants.

Listeners were tested separately or in pairs in a quiet experimental lab and were asked to perform two tasks sequentially. After presented with a statement, they first judged whether the statement conveyed some level of confidence by clicking YES or NO printed in two squares on the screen; they were informed that the level of confidence could be signaled by a lack of confidence or much confidence. If they answered YES, a 5-point scale was presented on the screen and they were asked to rate the speaker's level of confidence by choosing a number that best fit their impression of the last statement. A same number scale was also shown if they answered NO, however, they were asked to select any number to continue to the next trial.

2.2.3 Data analysis

A univariate analysis of variances were performed on the mean rating score of speaker's confidence for all statements of the two speakers, taking intended level of confidence (confident vs. close-to-confident vs. not confident), cue type (with lexical cue vs. no lexical cue), and type of utterance function (fact vs. intention vs. judgment) as three independent factors. Further analysis was planned when interaction between the factors were significant.

2.2.4 Results

Table 1 and Figure 1 demonstrate the mean perceptual ratings of with-cue and no-cue statements of each utterance type and in each level of confidence. The univariate ANOVA on all perceptual data revealed a significant main effect of level of confidence, $F(2, 1794)=2948.76$, $p<0.001$. Post-hoc Tukey's comparison confirmed that the speaker's confidence was rated the highest in the confident condition, followed by close-to-confident stimuli, followed by the unconfident stimuli. There was significant effect of cue type, $F(1, 1794)=215.78$, $p<0.001$, suggesting that with-cue statements were in general rated lower than the no-cue statements. The effect of type of utterance function was also significant, $F(2, 1794)=13.63$, $p<0.001$, suggesting that statements of intention were rated lower overall than statements of fact and judgment, with the

latter two not different from each other. Moreover, there was a significant interaction between level of confidence and cue type, $F(2, 1794)=136.20$, $p<0.001$ and a significant interaction between level of confidence and type of utterance function, $F(4, 1794)=2.40$, $p<0.05$. Separate analysis on each level of confidence revealed a significant effect of cue type for all levels of confidence: for confident, $F(1, 598)=51.80$, $p<0.001$, for close-to-confident, $F(1, 598)=257.28$, $p<0.001$, and for unconfident, $F(1, 598)=129.94$, $p<0.001$. Post-hoc comparison revealed a higher rating for the with-cue than the no-cue statement in the confident condition, whereas *lower* ratings were observed for the with-cue than no-cue statement in the close-to-confident and unconfident conditions.

Separate analysis also revealed an effect of type of utterance function for confident, $F(2, 598)=4.74$, $p<0.01$, for close-to-confident level, $F(2, 598)=12.98$, $p<0.001$, but not for unconfident level, $F(2, 598)=1.70$, $p>0.1$. These findings suggested that the lower ratings for intentions than for the other two functional types were most prominent in the confident and the close-to-confident conditions.

In order to examine the difference between the intended-neutral and the other levels of confidence, an additional ANOVA on the no-cue statements only revealed a significant effect of level of confidence, $F(3, 1195)=670.58$, $p<0.001$. Post-hoc comparison revealed that the intended-neutral utterances were rated as comparable to the close-to-confident utterances, lower than confident and higher than the unconfident utterances.

Table 1. Mean and standard deviation of the rating score of speaker's level of confidence for with-cue and no-cue statements in each intended confidence level.

	Confident				Close-to-confident			
	With Cue		No Cue		With Cue		No Cue	
Utterance Type	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Fact	4.54	0.29	4.36	0.30	3.31	0.38	4.08	0.36
Intention	4.50	0.29	4.28	0.31	3.28	0.47	3.75	0.61
Judgment	4.56	0.27	4.42	0.29	3.48	0.37	4.00	0.48

	Unconfident				Neutral	
	With Cue		No Cue		No Cue	
Utterance Type	Mean	SD	Mean	SD	Mean	SD
Fact	2.29	0.36	2.84	0.66	4.14	0.33
Intention	2.24	0.33	2.70	0.59	3.91	0.32
Judgment	2.29	0.37	2.76	0.65	4.05	0.31

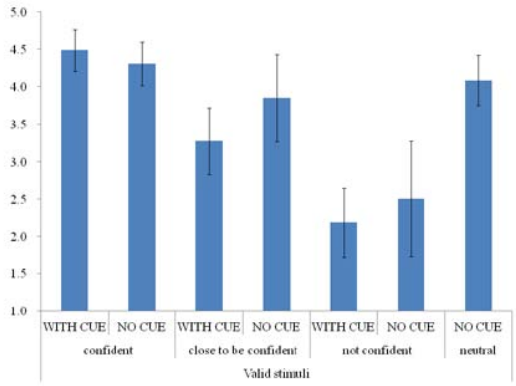

Figure 1. Perceptual rating scores of speaker's confidence for both with-cue and no-cue statements in each intended level of confidence

2.3 Acoustic study

2.3.1 Data analysis

In order to evaluate how different levels of confidence were acoustically differentiated in statements perceived as conveying confidence without lexical cues, acoustic measures were derived and analyzed only for "no-cue" statements. Due to the large number of intended neutral exemplars perceived as conveying some level of confidence, the statements with the frequency of "yes" response below 6 out 10 in the binary task were assigned as neutral, the statements with the frequency of "yes" response above 8 out of 10 were assigned as with-confidence. Among the with-confidence statements, the target level of confidence was re-assigned based on the perceptual results (see below). Those with the mean perceptual score above 4.2 in the 5-point rating task were designated as "confident", those with a mean score between 3.2 and 3.8 were designated as sounding "close-to-confident", and statements with a mean score below 2.8 were assigned to the "unconfident" condition.

Five acoustic measures that frequently differentiate among vocal emotion categories [2] [3] [4] were analyzed, including the mean fundamental frequency (mean f0, in Hertz), the range of fundamental frequency range (f0 variance, in Hertz), the mean amplitude (mean amplitude), the range of amplitude (amplitude variance), and speaking rate (in syllables per second). A normalization procedure was applied to the first four measures before comparing between speakers [2] [4]. Acoustic analyses were performed using Praat speech analysis software; the results for statements of fact for one male and one female speaker are reported here.

A multivariate analysis of variance (MANOVA) was performed on all valid no-cue statements that fell in the perceptual range defined for each target level of confidence. The mean and range of f0, mean and range of amplitude and speech rate were taken as dependent factors and the target level of confidence was considered as independent factors. A series of univariate ANOVA were further performed on each acoustic measure. We report the perceptual data for no-cue

statements from two speakers (1 male and 1 female) in the following section.

2.3.2 Results

Based on preliminary analysis of the two speakers, a total of 192 recordings were subjected to analysis. Table 2 demonstrated the mean perceived level of confidence and the values for each of the five acoustic parameters computed for the no-cue statements that were reassigned based on the ranking of the confidence perception. The one-way MANOVA was performed on acoustic data with the four perceived levels of confidence as independent variables and seven acoustic features (normalized mean f0, normalized f0 range, normalized mean amplitude, normalized amplitude range and speech rate) as dependent variables. The MANOVA indicated that the effect of level of confidence on the linear combination of the five acoustic parameters was significant, Wilk's λ = 0.575, $F(21, 523)=5.30$, $p<0.001$. Subsequent univariate analysis on each acoustic parameter revealed that the effect of level of confidence was significant for mean f0, $F(3, 188)=36.93$, $p<0.001$, f0 range, $F(3, 188)=4.16$, $p<0.01$, mean amplitude, $F(3, 188)=3.34$, $p<0.05$, amplitude range, $F(3, 188)=2.96$, $p<0.05$, and speech rate, $F(3, 188)=2.49$, $0.05<p<0.1$. Post hoc (Tukey's) comparisons revealed that the normalized mean f0 increased in value over neutral, confident, close-to-confident and unconfident meanings; the neutral level revealed a higher normalized f0 range and a higher speech rate than the three with-confidence levels, which did not exhibit any differences. Statements perceived as confident displayed both a higher mean amplitude and a higher amplitude range than statements in the other three conditions, which did not show any differences in amplitude measures. Statements perceived as not conveying any confidence were spoken more quickly overall than statements conveying the three levels of confidence.

Table 2. Mean normalized acoustic values for the no-cue statements produced by two speakers to express three levels of confidence and to the neutral level based on the perceived level of confidence.

Perceived level of confidence	Confidence rating	Mean f0	f0 variation	Mean amplitude	Amplitude range	Speech rate
Confident	4.50	0.40	1.37	1.23	1.96	4.61
Close-to-confident	3.60	0.65	1.86	0.96	1.58	4.50
Unconfident	2.30	0.76	1.85	1.02	1.64	4.27
Neutral	3.90	0.28	1.09	1.14	1.80	5.05

3. Discussion

This study demonstrates that listeners make use of prosodic information in speech to perceive different levels of confidence of a speaker when expressing facts, intentions, and making judgments which occur frequently in discourse. In particular, statements rated as conveying different levels of confidence were acoustically distinct in their f0 characteristics, with the more confident the statement intended to be the higher the confidence rating, among other acoustic distinctions. The additional presence of a lexical phrase that signaled speaker confidence had varying effects on perceptual ratings depending on the level of speaker confidence communicated: the presence of the lexical cue in confident statements tended to amplify confidence ratings, whereas lexical phrases in statements that lacked full confidence (close-to-confident, unconfident) tended to attenuate impressions of confidence, yielding lower ratings for "with-cue" utterances in these conditions. Interestingly, different utterance types (facts, intentions, judgments) differed in subtle but significant ways in how listeners perceived speaker confidence from prosodic cues. These findings are broadly consistent with Pell [5] who found that the attitudinal and interpersonal significance of prosody can be perceptually differentiated by both healthy and right-hemisphere-damaged adults, extending these findings to a group of healthy young adults.

The acoustic analysis on the no-cue statements further demonstrated several important mechanisms underlying the acoustic realization of confidence-related information in speech: 1) the f0 variance and the speech rate differentiated the with-confidence and neutral statements; 2) the mean f0 predicted the levels of confidence in a parametric way; and 3) the mean amplitude and the amplitude change highlighted the confident-level specifically. Taken together, our findings indicate that prosodic information guides how listeners infer the confidence state of a speaker, and that these pragmatic inferences may be affected by both linguistic and paralinguistic cues in speech communication, and may vary for different speech acts.

4. References

[1] Scherer, K., London, H. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality, 7,* 31-44.

[2] Pell, M.D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S.A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37,* 417-435.

[3] Cheang, H.S. & Pell, M.D. (2008). The sound of sarcasm. *Speech Communication, 50,* 366-381.

[4] Liu, P., & Pell, M.D. (2012). Recognizing vocal emotions in Mandarin Chinese: A validated database of Chinese vocal stimuli. *Behavior Research Methods, 44,* 1042-1051.

[5] Pell, M.D. (2007). Reduced sensitivity to prosodic attitudes in adults with focal right hemisphere brain damage. *Brain and Language, 101,* 64-79.

5. Acknowledgements

This study was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada to Marc D. Pell. We are very grateful to Sonia Kroll, Lorraine Chuen, Nabitha Kanagaratnam, Mary Giffen, Caleb Harrison, and François Anderson for their assistance in the study.