Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 51 (2009) 640–655

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Charisma perception from text and speech

Andrew Rosenberg *, Julia Hirschberg

*Columbia University, 2960 Broadway, New York, NY 10027-6902, United States*

Received 3 August 2006; received in revised form 7 November 2008; accepted 10 November 2008

**Abstract**

Charisma, the ability to attract and retain followers without benefit of formal authority, is more difficult to define than to identify. While we each seem able to identify charismatic individuals – and *non*-charismatic individuals – it is not clear what it is about an individual that influences our judgment. This paper describes the results of experiments designed to discover potential correlates of such judgments, in *what* speakers say and the *way* that they say it. We present results of two parallel experiments in which subjective judgments of charisma in spoken and in transcribed American political speech were analyzed with respect to the acoustic and prosodic (where applicable) and lexico-syntactic characteristics of the speech being assessed. While we find that there is considerable disagreement among subjects on how the speakers of each token are ranked, we also find that subjects appear to share a functional definition of charisma, in terms of other personal characteristics we asked them to rank speakers by. We also find certain acoustic, prosodic, and lexico-syntactic characteristics that correlate significantly with perceptions of charisma. Finally, by comparing the responses to spoken vs. transcribed stimuli, we attempt to distinguish between the contributions of "what is said" and "how it is said" with respect to charisma judgments.
© 2008 Elsevier B.V. All rights reserved.

*Keywords:* Charismatic speech; Paralinguistic analysis; Political speech

## 1. Introduction

Charismatic individuals have been defined as those who command authority by virtue of their personal qualities rather than by formal institutional or military power (Weber, 1947). How they acquire authority, however, is a question of considerable discussion. While some see charisma arising primarily from the faith of a leader's *listener–followers* (Marcus, 1967), others believe that it arises from particular individual's *gift of grace* and *an inspiring message*, and triggered by an *important crisis* (Boss, 1976). However, all who study charisma concur in believing that charismatic leaders share a particular ability to communicate. Leaders widely believed to be charismatic, such as Martin Luther King Jr., Fidel Castro, Adolf Hitler, and Pope John Paul II, are also particularly noted for their oratorical abilities.

In this paper, we investigate the language-based aspects of charisma. In particular we are interested in identifying aspects of *what* speakers say and *how* they say it as potential correlates of others' judgments of their charisma, or lack thereof. We describe two perception experiments designed to identify possible acoustic, prosodic, and lexico-syntactic characteristics of charisma, one using spoken data and the other using transcribed and written materials from the same speakers. We correlate subjects' judgments of this material with lexico-syntactic and acoustic and prosodic features of the assessed speech and with lexico-syntactic characteristics of the text tokens. Finally, we compare judgments from text transcriptions alone to judgments made from speech, to distinguish the contributions of how something is said from what is said in subject judgments of charisma.

Our motivation for this study is two-fold: on a scientific level, we are interested in determining whether speakers who are judged charismatic share certain acoustic and prosodic characteristics, and how these interact with lexical content and syntactic form. While communicative talent

* Corresponding author. Tel.: +1 646 567 7747.
*E-mail addresses:* amaxwell@cs.columbia.edu (A. Rosenberg), julia@cs.columbia.edu (J. Hirschberg).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.11.001

has been widely assumed in the literature on charisma to contribute to the charismatic appeal of an individual, there is no theoretical framework on the role that the form and content of charismatic individual's speech or writings plays in overall charisma judgments. However, most of the specific features we have tested derive from claims or speculations in the literature about different characteristics of charismatic speech.

From a technological point of view, we believe that such research has potential applications in speech synthesis and speech understanding: first, a better understanding of the acoustic and prosodic characteristics of charisma in human speech could support the generation of more 'charismatic' synthetic speech for applications intended to be persuasive and compelling, such as commercial and political advertisements or telephone solicitations. Second, such an understanding might support the automatic identification of 'charismatic' speakers, who in turn are likely to be successful in attracting a political, military, or religious following. Finally, knowledge of how charismatic individuals speak has the potention to support the creation of online training systems that help individuals to become more charismatic speakers themselves.

In Section 2, we discuss previous research on charisma, particularly with respect to charismatic language, in the sociology, rhetoric, and natural language processing literature. In Section 3, we describe an online experiment we conducted to elicit subject judgments of charisma and other personal attributes of speakers of tokens of public speech. Section 4 describes a parallel experiment in which similar judgments were elicited from subjects based on transcripts of the spoken tokens described in 3. We conclude in Section 5 and describe future research in Section 6.

## 2. Related research

In this investigation into the spoken and lexico-syntactic aspects of charismatic speech, we were guided in the design of our experiments and in many of our initial experimental hypotheses by the previous work of sociologists and rhetoricians. Following Weber's (1947) discussion of CHARISMATIC AUTHORITY as a legitimate source of leadership, social scientists have worked toward defining what exactly charisma is. Marcus (1967) argued that faith in the leader was necessary to charisma citing Adolph Hitler as an example. "[T]he 'true-believing' Nazi had implicit faith that under the Fuhrer's leadership Germany could master the destiny of history..." (p. 237). Boss (1976) identified a set of 'essential aspects of charisma' which he felt were directly related to the rhetoric employed by a potential leader. These nine aspects included "(1) the 'gift of grace' ...; (2) the concept of the 'leader–communicator'; (3) the 'inspiring message'; (4) the 'idolatrous follower'; (5) a shared history; (6) high status; (7) the concept of 'mission'; (8) an important crisis; and (9) successful ...results" (p. 301). Most relevant to our study is (3), the 'inspiring message', although what makes a message 'inspirational' – either in form or in content – is little discussed. More recently, Bird (1993) has explored the role of charisma in the propagation of NEW RELIGIOUS MOVEMENTS (NRMs), more popularly known as 'cults', finding that NRM leadership "has almost wholly assumed a personal charismatic form".

A number of authors have examined the role of communicative skills in defining a charismatic or persuasive leader in more depth. For example, Hamilton and Stewart (1993) propose an information processing model of persuasion. They describe an experiment in which subjects were presented with a set of messages based on a template concerning the dangers of excessive exercise (p. 239) and asked to evaluate the dynamism, competence and trust of the speaker of the message. The experimenters manipulated the language intensity of the message by including 'high', 'moderate' or 'low' intensity lexical items within a template text. 'High' intensity words contain more emotional content and/or are more specific than 'low' intensity words. The experimenters, using a causal modeling program, observed that when subjects perceived a message as more intense they found the source of the message to be more dynamic. Sources perceived as highly dynamic were also perceived as highly competent, and competent speakers were perceived as more trustworthy. They proceed to describe this interaction between dynamism, competence and trust as 'the *charisma sequence*'.

Touati (1993) examines charisma in the context of French political speech, comparing the prosody of politicians before and after an election. He claims that this comparison captures the transition "when persuasion (when a politician aims to gain votes) gives [way] to objective pathos (when a politician comments [on] his political victory or defeat)" (p. 168). He finds that pre-electoral, 'persuasive' speech is characterized by an increased pitch range and variation of pitch register compared to post-electoral speech. Thus, particular prosodic features are associated with attempts to project charisma. We include these among the prosodic features we test in our perception studies.

Persuasive speech has also been described in terms of its rhetorical structure and the coherence of its arguments (e.g. Cohen, 1987). It is very likely that the ability to persuade may be an important attribute of those identified as charismatic. However, theoretical research on charisma claims that charismatic leaders have something more – the ability to develop an 'intimate relationship' with listeners or readers, involving trust and 'an inspiring message' in addition to a persuasive argument (Boss, 1976). While there has been relatively little experimental or quantitative work in this area, Tuppen (1974) reports an interesting experiment on a related topic, attempting to quantify communicator credibility. In this experiment, subjects were asked to read short character sketches of 10 communicators, and rate each of them in a terms of 64 personal attributes, 28 using bipolar adjective scales (e.g. Honest–Dishonest, Bold–Timid) and 36 using seven-point Likert scales (e.g. "I can trust the judgment of the speaker", "I should like to have the speaker as a personal friend"). Subject ratings were subsequently clustered and the dominant ratings were used to define the cluster. Tuppen independently assigns the label ‘charisma’ to a cluster defined by the following adjectives: “convincing, reasonable, right, logical, believable, intelligent; whose opinion is respected, whose background is admired, and in whom the reader has confidence”. We have tested a number of Tuppen’s attribute scales in our own perception studies here.

While previous studies of charisma have postulated a number of factors which might play a role in an individual’s perception as charismatic, and while both the form and content of communication has been assumed to be an important aspect of perceived charisma, there has been very little empirical research on what particular aspects of language contribute to perceptions of charisma and, in particular, of the relationship of acoustic, prosodic, and lexico-syntactic form to content in producing this effect. Our work attempts to address these questions by an empirical investigation of speech judged charismatic and non-charismatic. In the course of our studies, we have tested many features proposed in the literature as well as some new features we propose here.

## 3. Charisma judgments from speech

### 3.1. Experimental design

In order to look for objective correlates of charisma in individuals’ spoken and written productions, we first needed to address some basic issues about charisma perception and potential confounds. Would subjects we asked to provide judgments share a common definition of charisma? Would they apply this term similarly to a given set of speech productions? If not, would it be possible to construct a ‘functional’ definition of charisma from other attributes subjects might be able to rate productions on more easily? With respect to the materials subjects would be asked to rate, would judgments be influenced by the identity of the item’s speaker, by the topic of the token, or by the genre or style of the token? Our experimental design attempted to address each of these possible concerns.

First, we chose our materials to balance speakers, topics, and genres. We looked for speech tokens from a small set of speakers, whose public speech covered a similar set of topics, and for whom speech tokens could be found in a wide variety of genres, or speaking styles. Since the experiment was designed during the winter and spring of 2004, we found abundant speech material available for the nine candidates running at that time for the Democratic Party’s nomination for President: Sen. John Kerry, Rep. John Edwards, Gov. Howard Dean, Rep. Richard Gephardt, Rev. Al Sharpton, Amb. Carol Moseley Braun, Rep. Dennis Kucinich, Gen. Richard Clark, and Sen. Joseph Lieberman. We chose speakers from the political field for a number of reasons. We hypothesized that at least some of these politicians would demonstrate charismatic qualities in their speech. Also, the varied activities of the candidates ensured that speech would be available from different genres: interviews, debates, stump speeches, and campaign ads. We limited our speakers to Democrats to confine the range of opinions presented in the tokens, as it has been suggested in the literature (Boss, 1976; Dowis, 2000; Weber, 1947) that a listener’s agreement with a speaker bears on their judgment of the speaker’s charisma. We selected segments from a variety of topics to test the influence of topic on subject judgments of charisma. We included five speech tokens from each speaker, one on each of the following topics: health care, postwar Iraq, Pres. Bush’s tax plan, the candidate’s reason for running, and a content-neutral topic (e.g., greetings). For these five tokens, we varied genre among the following types: interview, debate, stump speech, campaign ad. Since the speech tokens came from a variety of sources and recording conditions, we normalized the tokens for intensity to $-12$ dBFS.

From a large set of segments which fit the above criteria, we then screened potential tokens to judge whether a token ‘sounded charismatic’ or not. This rough evaluation was intended to balance the ‘charismatic’ and ‘non-charismatic’ tokens across speakers and topics.[1] In total, 22 of the 45 tokens used in the experiment were judged ‘charismatic’ by the experimenters. Tokens varied in length from 2 to 28 s. Given the other constraints we imposed, balancing by topic, speaker and whether or not collaborators found the material to be charismatic, identifying tokens of a consistent length proved difficult. The mean token length was 10.09 s, with a standard deviation of 4.98 s. The topic which caused the greatest difficulty was the content-neutral topic. These were commonly short greetings, for example “It’s a pleasure to be with you today”. The mean length of these tokens was 4.65 s. When possible each token contained a single sentence. However, we only spliced out tokens at silent regions leading two tokens that were longer than 16 s. Due to experimental error, one of the speech segments was presented twice (Rep. Edwards’ ‘reason for running’), and another omitted (the content-neutral statement from Rep. Gephardt). While this error skewed the balance of the corpus as a whole slightly, it also allowed us to do a post hoc test for intra-rater consistency.

To determine whether subjects shared a common definition of charisma, we asked them to rate tokens of political speech with respect to the statement ‘The speaker is charismatic’ on a five-point Likert scale, from ‘disagree completely’ (1) to ‘agree completely’ (5). Below, we term this statement ‘the charismatic statement’. To determine whether subjects shared a functional definition of charisma, independent of their intuitive definition of the term, we also asked them to rate token speakers on 23 additional

[1] Judd Sheinholtz, Aron Wahl, and Svetlana Stenchikova participated in the selection and screening of tokens, together with the authors. Segments that we could not agree on, or considered to be only moderately charismatic, were not included in the materials.

attributes drawn from claims and findings in the previous literature on charisma (many, in particular, from Tuppen (1974), on the same Likert scale). These statements were of the form "The speaker is X", where X was one of the following: charismatic, angry, spontaneous, passionate, desperate, confident, accusatory, boring, threatening, informative, intense, enthusiastic, persuasive, charming, powerful, ordinary, tough, friendly, knowledgeable, trustworthy, intelligent, believable, convincing, and reasonable. These attributes represented a subset of those proposed in the literature as positively or negatively associated with charisma. We also included "The speaker's message is clear" and "I agree with the speaker" as statements to be rated, again based upon hypotheses in the literature about the sources of charisma. The full set of statements as presented to the subjects is shown in Appendix A.

The experiment was administered via the internet between May 18 and June 3, 2004. The subjects for the study were eight native speakers of standard American English with no reported hearing problems, recruited via email and over the internet. They were presented with the experimental materials via a standard web browser. Each token was played simultaneously with the presentation of a web form, illustrated in Appendix A. The clip was repeated with two seconds of silence between iterations until the subject had responded to all 26 statements, and had moved on to the next segment. The order of presentation of the 45 tokens was randomized for each subject. Additionally, the order of the 26 statements was randomized for each token. No restriction was placed on the order to which the statements could be rated. At the completion of the survey, each subject was asked to identify by name any speakers they felt they had recognized at any point during the presentation of stimuli. This gave us a rough approximation of whether subject belief in the recognition of a speaker might have affected judgments of charisma or other attributes without encouraging the subjects to focus on the identity of the speaker while responding to the statements.

It took users an average of 1.5 h to complete the survey. The shortest time taken by any subject was 49.5 minutes; the longest ~3 h. We note that the long duration of the study contributed to the fact that only a relatively small number of subjects completely the survey, and this, in turn, limits the conclusions we may be able to draw from the experiment. However, the results discussed below do provide some empirical data on previous hypotheses as well as indicating some fruitful areas for further testing on a larger scale.

### 3.2. Speech study analysis and results

The analysis of subject judgments on our spoken corpus consisted of four parts: we first examined overall subject agreement on ratings for all speech tokens and all of the statements about these tokens that subjects were asked to rate, to determine how consistent these ratings were. We also wanted to determine whether subjects shared a common 'functional' definition of charisma, in terms of the additional characteristics they attributed to statements they found more or less charismatic. We then examined the potential influence of other factors on charisma judgments, including order of presentation, the topic and genre of particular tokens, and the identity of a token's speaker. Finally, we looked for possible correlations between subjects' charisma ratings for tokens and the acoustic, prosodic, and lexico-syntactic features of those tokens, to address our main question: what makes charismatic speech charismatic.

#### 3.2.1. Across-subject agreement on ratings

To examine overall inter-subject agreement on ratings for all tokens and statements, we used the weighted kappa statistic (Cohen, 1968) with quadratic weighting.[2] The mean $\kappa$ value over all 45 tokens and 26 statements was 0.213. This is rather low agreement and indicates a fair amount of individual variation in the ratings of at least some of the 26 statements or some of the tokens. In order to identify potential sources for this variation, the kappa contribution from each of the 45 tokens was examined individually. This breakdown allowed us to determine which of the tokens were most and least consistently ranked across subjects. Similarly, we computed the kappa contribution from ratings of each of the 26 statements.

We found no significant differences in kappa values across the particular tokens used in the experiment. Subjects did *not* exhibit greater or less consistency for any of the particular speech segments they heard. However, tokens spoken by Rep. Edwards and Sen. Lieberman show significantly (ANOVA $p = 5.4 \times 10^{-21}$) *more* agreement across all statements than tokens spoken by the other seven. Interestingly, these two speakers are rated as the most and least charismatic speakers of the group of nine. There was, moreover, a substantial range of inter-rater agreement with respect to the 26 individual statements subjects were asked to rate for each token. Of particular note is the contrast between the statements that showed the greatest and least agreement. Tables 1 and 2 contain the five statements with the highest and lowest kappa scores, respectively.

Statements corresponding to dynamic, high-activation emotions (accusativeness, passion, intensity, anger, and enthusiasm) ranked among those most consistently rated. However, agreement on ratings of trust, reasonability, believability, desperation, and ordinariness rank hardly greater than what would be expected by chance. This might have arisen from subjective differences with respect to perceptions of qualities such as trustworthiness or believability. Alternately, subjects may have been skeptical of political speech, and therefore reluctant to ascribe qualities

[2] While, kappa is not the only way to determine the correlation between responses, it is widely used in computational linguistics and produces easily understood results.

Table 1
Statements with most consistent inter-subject agreement in speech survey.

| Statement | $\kappa$ |
|---|---|
| The speaker is accusatory | 0.512 |
| The speaker is passionate | 0.458 |
| The speaker is intense | 0.431 |
| The speaker is angry | 0.404 |
| The speaker is enthusiastic | 0.362 |

Table 2
Statements with least consistent inter-subject agreement in speech survey.

| Statement | $\kappa$ |
|---|---|
| The speaker is trustworthy | 0.037 |
| The speaker is reasonable | 0.070 |
| The speaker is believable | 0.074 |
| The speaker is desperate | 0.076 |
| The speaker is ordinary | 0.115 |

such as ‘being reasonable’ to politicians, while emotions such as anger and enthusiasm may be less evaluative.

Ratings of the charismatic statement yielded a mean kappa score of 0.224. While this kappa value is low – considering that a value of 0 indicates agreement equal to chance, and 1 indicating perfect agreement – it is important to recognize that the evaluated qualities are subjective. Here, and elsewhere in the results, the metrics of agreement conflate two factors: the degree of similarity between subjects' underlying concepts, and the degree of similarity in their responses to particular stimuli. For example, the kappa statistic evaluates whether two subjects mean the same thing by ‘charisma’ *and* whether they observe the same degree of charisma in the same tokens. A kappa value of 0.224 places “The speaker is charismatic” as the eighth most consistently labeled statement. Despite this low agreement, it is of note that subjects agreed about charisma more than about such qualities as intelligence ($\kappa = 0.119$) (“The speaker is intelligent”) and confidence ($\kappa = 0.215$) (“The speaker is confident”).

### 3.2.2. Within-subject correlation of statement ratings

To see whether subjects agreed upon a common ‘functional’ definition of charisma in terms of other speaker attributes we asked them to judge, we next examined which statement ratings were positively or negatively correlated with ratings of charisma. We again applied Cohen's kappa statistic with quadratic weighting to determine the correlation between the charismatic statement and the remaining 25.[3] Those statements that demonstrated the greatest positive or negative correlation with the charismatic statement appear in Table 3. We conclude that our subjects' ‘functional’ definition of a charismatic speaker is ‘one who is enthusiastic, charming, persuasive, passionate, convincing – and *not* boring’. Our findings support Dowis's (2000) and Boss's (1976) claims that enthusiasm and passion are positively correlated with charisma, while boringness is negatively correlated.

Table 3
Statements showing most consistent positive and negative correlation with charismatic statement as determined by mean $\kappa$ scores.

| Statement | Mean $\kappa$ |
|---|---|
| The speaker is enthusiastic | 0.606 |
| The speaker is charming | 0.602 |
| The speaker is persuasive | 0.561 |
| The speaker is boring | −0.513 |
| The speaker is passionate | 0.512 |
| The speaker is convincing | 0.503 |

Note that ratings of the desperate, threatening, accusatory, and angry statements showed neither a positive nor negative ($|\kappa| < 0.15$) correlation with the charismatic statement. It is particularly interesting that ratings of a speaker's anger (shown to be relatively consistently rated across subjects in Section 3.2.1) had no impact in either direction on a subject's judgment of the speaker's charisma. Since anger is a polarizing, high-activation emotion, we hypothesized that it would show some positive or negative correlation with charisma, possibly with an interaction with subject reports of agreement (or lack thereof) with the speaker. However, we observe neither of these.

### 3.2.3. Influence of speaker, topic, genre, and order of presentation on charisma ratings

For our subjects, the speaker of a segment significantly influenced ratings of charisma ($p = 1.75 \times 10^{-10}$).[4] Mean ratings for each speaker indicate the following ordering, from most to least charismatic: Rep. Edwards (mean rating 3.75), Rev. Sharpton (3.40), and Gov. Dean (3.33). The three least charismatic were Sen. Lieberman (2.38), Rep. Kucinich (2.73), and Rep. Gephardt (2.77). As determined by our ‘exit’ survey of whether subjects believed they had recognized any of the speakers (described in Section 3.1), the mean number of speakers claimed to have been recognized by subjects was 3.25 (out of the 9 speakers) with a maximum of 6 and a minimum of 0. Subjects rated tokens spoken by a (claimed) recognized speaker as more charismatic (mean rating 3.28) than those spoken by unrecognized speakers (mean rating 2.99). This difference was significant with $p = 0.007$. This finding may suggest either that familiarity with a speaker positively influenced perceptions of charisma – or that charismatic speakers are more recognizable than uncharismatic speakers.

The genre in which the speech token was delivered also significantly influenced subject ratings of charisma ($p = 0.0058$). Speakers were rated as more charismatic when they were delivering a stump speech (mean rating 3.28) than when they are being interviewed (2.90). Speech segments extracted from debates (3.10) were rated in line with the

[3] We note that the cardinality of the correlations was consistent whether measured using Cohen's kappa or Spearman's $r$ statistics.

[4] All $p$ values in Section 3.2.3 are determined by one-way ANOVA with repeated measures.

overall mean (3.09) with respect to charisma. The corpus contained only one segment that was taken from a campaign advertisement; while this segment was rated as below average in charisma (2.88), this obviously cannot be taken as reflective of the genre as a whole. The impact of genre on subject ratings may be easily explained: the enthusiasm and dynamism that can be appropriately conveyed during a stump speech – at least, by speakers who *can* convey charisma – may be less appropriate in an interview.

The topic of the segments used in our experiment (post-war Iraq, health care, taxes, reason for running, and content-neutral) had no statistically significant impact on subjects' ratings of charisma. While the semantic content of a particular speech segment may contribute to the perceived charisma of its speaker, the general topics we analyzed do not appear either to promote or to inhibit perceptions of charisma.

As noted in Section 3.1, due to experimental error, one of our speech tokens was presented to subjects twice. So we were able to compare subject ratings on the two different presentations of the same token to measure consistency. While no subject ratings varied significantly between presentations (mean difference of 0.4), ratings of the tough, ordinary and charismatic statements varied most. The mean charisma rating of tokens was greater on the second presentation by 0.43 a difference that is not statistically significant. Further study, particularly of charisma judgments, will be needed to determine if this variation is meaningful.

### 3.3. Token features associated with charisma

As described in Section 3.2.1, subjects' agreement on the charisma of individual tokens was modest. However, we are able to assess study-wide interactions between ratings of charisma and a variety of lexical, syntactic, and acoustic–prosodic properties, across all subjects. The properties we examined were chosen based on claims and findings in the previous literature as well as our own intuitions and previous findings in earlier studies of other types of speaker characterization. The analysis below relies on Pearson's $R$ test to measure correlation between numeric properties and charisma ratings and ANOVA to model the interaction between categorical properties and the assessment of charisma. The null hypothesis when applying each of these statistical instruments is "there is no interaction between the property and ratings of charisma". Due to the low agreement we observed in the ratings of charisma, we did not aggregate these ratings in any way before performing the analyses discussed below. That is, for each token there are eight charisma ratings, one from each subject, included in the analysis of any property. Thus, this lack of aggregation allows us to reject the null hypothesis at a study-wide level. That is to say, while two subjects may differ to what degree they observe charisma in particular tokens, they may respond similarly to lexico-syntactic or acoustic/prosodic properties in terms of the way these properties correlate with their perception of charisma.

#### 3.3.1. Lexico-syntactic properties of charismatic speech

We examined a number of lexico-syntactic features of the tokens presented in the perception study to see how these correlated with subjects' ratings of the charisma of the token. Most of these features were chosen to test claims in the previous literature, either directly or by approximating a more general claim, such as the 'simplicity' of a charismatic speaker's message. Our lexico-syntactic features included: the number of words in the token, the ratio of function to content words in the token, the number of repeated words, a measure of lexical complexity due to Dowis (2000), the token's pronoun 'density', and the ratio of disfluencies to number of words in the token.

We first looked at the amount of spoken material in each token, as determined by length in words, to test whether the sheer amount of a speaker's speech influenced charisma judgments. Charismatic leaders such as Castro and Hitler were famous for their lengthy speeches, suggesting that, on a more local measure, duration of token would be useful to examine. In fact, we found that the number of words in the token positively correlated with judgments of charisma at a rate approaching significance with $r = 0.097$ and $p = 0.068$. The more material presented to the subject, the more charismatic the speaker was perceived to be.

Following Hamilton and Stewart's information-centric view of charisma (Hamilton and Stewart, 1993), we also hypothesized that, the more relative content there is in a message, the more likely it is that such content can influence the charisma rating. To quantify content to a first approximation, we used a simple metric — the ratio of function (prepositions, pronouns, determiners, conjunctions, modal verbs, auxiliary verbs, and particles) to content words (e.g. nouns, verbs) in each token. We tagged the part of speech of the words in each token automatically using the Brill tagger (Brill, 1995), and calculated this ratio using the resulting part-of-speech tags. This measure also approached significant correlation with ratings of charisma ($r = 0.102$, $p = 0.0569$), suggesting that, the fewer content words relative to the number of function words – the *less* 'content' in the message – the more charismatic the speaker is perceived to be. A related measure of content might be the presentation of *new* content. So we examined the number of repeated words in each token to see how repetition or its lack might correlate with perceptions of charisma. Consistent with our findings for a lack of 'content' in charismatic speech, we found that the lack of new content – or, repetitions' *positively* influenced judgments of charisma with $r = 0.0986$ and $p = 0.0645$, approaching significance. Repeating oneself of course is a common rhetorical device, whether to "drive a point home" or in employing anaphoric expressions. Further analysis of the syntactic structure of these relatively 'content-free' tokens will be needed to see whether this result can be attributable to particular rhetorical form.

Our findings with respect to lack of content words and use of repetition run somewhat counter to the general tenor of the existing charisma literature, which places much importance on the content of a charismatic speaker's message. However, it is more consonant with another widely held believe, that a charismatic speaker's message is 'simple'. Dowis (2000) posits that simpler words are more effective than complex terms in delivering a charismatic message. He proposes a simple measure of the complexity of a lexical item – the number of syllables it has. However, when we computed the number of syllables per word for each token, we found that this metric influenced ratings of charisma in the opposite direction to that predicted by Dowis. Greater mean syllables per word corresponded to higher ratings of charisma; or, more 'complex' words characterize charismatic speech. This influence was significant with $r = 0.123$ and $p = 0.021$. We hesitate to generalize too broadly here, but our findings present at least some empirical counter-evidence to Dowis' anecdotal claims. Taken together with our findings above on the lack of 'content' in charismatic speech, we might propose in fact that a different measure of language simplicity in terms of a message's content rather than its morphological structure might capture this intuition more effectively.

Charisma is often said to be a personal quality of a speaker, and a manifestation of a special *speaker–listener* relationship. The literature on charisma (Weber, 1947; Marcus, 1967; Boss, 1976) suggests that charismatic individuals have an unusual ability to establish a personal relationship with their followers, even without face-to-face contact. Such terms as 'father figure' are often used about such leaders. We hypothesized then that the presence of first and second person pronouns might characterize charismatic speech. We thus examined density of pronouns (ratio of pronouns to total words) broken out by first, second, and third person as a possible correlate of charisma judgments. We found that only the density of first person pronouns significantly influenced subject ratings of charisma ($r = 0.116$, $p = 0.0294$). No other pronoun measures showed any significant influence. So, at least some aspect of 'personal' speech seems to be present in charismatic speech, although only a rather egocentric one.

A final characteristic proposed in the literature to characterize charismatic speakers is the clarity of their message (Tuppen, 1974). One objective and rather literal measure of clarity is the absence of disfluency in a speaker's utterances. We examined the ratio of disfluencies – defined here as the sum of both false starts and filled pauses – to total words in a token, as a measure of clarity. We found this ratio to be *negatively* correlated with subject ratings of charisma ($r = -0.124$, $p = 0.0204$); the greater the disfluencies in a token, the less likely it was to be rated as charismatic. While disfluency may be an indicator that an utterance is less clear, it may also be interpreted as representing a lack of speaker confidence in their message, which may in turn serve to undermine how "convincing" a speaker is. Recall that how convincing a speaker is perceived to be is itself significantly correlated in our study with subjects' ratings of a speaker's charisma (cf. Section 3.2.2). Thus, on either basis, disfluency appears to be a clear negative correlate of charisma judgments in our corpus.

### 3.3.2. Acoustic–prosodic properties of charismatic speech

We also examined certain acoustic and prosodic characteristics of the speech tokens used in our study. Some of these features, such as pitch measures, had been found in the literature to correlate with charisma or at least with persuasive speech (e.g. Touati, 1993) while others have been observed more informally in the speeches of charismatic leaders. We examined pitch, intensity, speaking rate, and durational features and then measured the degree of correlation between these features and subject ratings of the charismatic statement. These features were calculated over the entire speech token, in each case, using Praat speech analysis software (Boersma, 2001). We also examined certain properties of the intonational contours of component intonational phrases within the tokens (labeled by hand) and performed similar correlations, to test our own hypothesis that certain contours in Standard American English (SAE) appeared to us to be more 'charismatic'.

We hypothesized that Touati's findings of variation in pitch range and standard deviation might be associated more generally with perceptions of speaker engagement and expressiveness. We also hypothesized that louder messages might be perceived as more 'commanding' and that variation in speaking rate might also play an important rhetorical device. Recall that our lexico-syntactic results (Section 3.3.1) had already indicated that longer messages measured in words were positively correlated with charisma judgments, so we expected to find that longer utterances measured in other ways would be similarly perceived as more charismatic.

To test these hypotheses, we first examined (raw) mean, standard deviation, maximum, and minimum $f0$ for all male speakers. Our findings are consistent with Touati's and extend his findings to other aspects of pitch behavior. All of these $f0$ properties positively and significantly correlated with ratings of charisma (mean $f0$: $r = 0.252$, $p = 1.69 \times 10^{-6}$; standard deviation of $f0$ $r = 0.129$, $p = 8.65 \times 10^{-3}$; max $f0$ $r = 0.183$, $p = 5.36 \times 10^{-4}$; min $f0$ $r = 0.126$, $p = 0.0177$). For all features, the greater the value of the feature, the greater the perceived charisma. For example, the higher the mean $f0$ and standard deviation of $f0$ for a given token, the more its speaker was perceived as charismatic. The higher mean, max and min $f0$ ratings can be seen as indicators that, indeed, the more charismatic speaker speaks 'up' in their pitch range. The high standard deviations of pitch in a token may lead listeners to judge the speaker as more expressive. This in turn, may signal some of the other attributes that correlate highly with charisma, such as enthusiasm (cf. Section 3.2.2) and dynamism, predicted in the literature by Boss (1976) and Tuppen (1974) as a correlate of charisma.[5]

[5] When we normalized these features by calculating *z-scores* for all speakers, male and female, only the $z$-score of a token's mean $f0$ approached significant correlation (positively) with charisma ratings ($r = 0.104$, $p = 0.0504$) – not maximum or minimum $f0$. When a token was higher in the speaker's pitch range, it was rated more charismatic. Standard deviation of $f0$ over all speakers, male and female, was also a significant correlate of charisma with $r = 0.127$, $p = 3.57 \times 10^{-3}$.

We examined intensity as an indicator of perceived loudness, under the hypothesis that louder and more forceful messages might convey a more charismatic impression. However, since we had previously normalized all tokens for intensity due to the differences in recording conditions of our data (cf. Section 3.1), we can only examine mean and standard deviation in our tokens. We found that only mean intensity approached significant correlation with charisma judgments ($r = 0.0718$, $p = 0.0549$), with louder utterances positively correlated with charisma ratings, as we had predicted. Variation in original recording conditions of the original tokens made a fuller analysis of the contribution of intensity to charisma perceptions impossible for this experiment.

Hypothesizing that speaking rate or variation in rate might correlate with charisma judgments, based on our observation that effective speakers often vary their rate to good effect, we calculated speaking rate in syllables per second for each token and compared it to ratings of charisma. Contra our earlier hypothesis, we found no correlation of variation in rate within a single token to be correlated with charisma judgments. However, we found instead a *positive* correlation of rate itself with judged charisma, approaching significance, with $r = 0.094$ and $p = 0.0902$. Nothing we have found in the literature has suggested that faster speech is correlated with charisma. However, insofar as slower speech might convey hesitation and doubt, the finding might be explainable. Further experimentation will be required to determine the nature of the interaction between speaking rate and charisma.

To test our hypothesis that intonational contour might play a role in perceptions of charisma – that particular contours or perhaps particular phrasal ending patterns contributed to a speaker's charisma ratings – we manually labeled our tokens using the ToBI (Silverman et al., 2:867–870) scheme for SAE. We examined possible correlations between pitch accent type, phrase contour type and phrase boundary behavior and ratings of charisma. As each token could contain a number of pitch accents or intermediate phrases, the features we used for analysis were distributions of the available classes. We classified intonational phrase boundary patterns into three classes: rising ($L-H\%, H-H\%$), falling ($L-L\%, L-$), and plateau ($H-L\%, H-$).[6] We found that an increase in a token's proportion of rising phrase boundaries negatively correlates with charisma ($r = -0.172$, $p = 0.00119$). Rising phrase final behavior may signal that the phrase contains a FORWARD-REFERENCE (Pierrehumbert and Hirschberg, 1990) and often can be observed in questions or in speaker uncertainty, neither of which seem plausibly associated with 'persuasiveness' or 'convincing' behavior, which themselves are highly correlated with charisma.

We also found some interesting correlations, both positive and negative, for pitch accent type and ratings of charisma. A greater proportion of $H^*$ pitch accents, the most common pitch accent type in SAE, positively correlated with ratings of charisma ($r = 0.145$, $p = 0.00658$). This accent is commonly associated with the presentation of 'new' information. We also found that the presence of greater proportions of $L^* + H$ ($r = -0.111$, $p = 0.0363$) and of $L^*$ ($r = -0.223$, $p = 2.28 \times 10^{-5}$) pitch accents were negatively correlated with ratings of charisma. $L^* + H$ accents are typically used to convey 'uncertainty' or 'incredulity' (Pierrehumbert and Hirschberg, 1990) while L* accents are used to present information that is known or inferrable among discourse participants – *given*, or 'old' information. These findings might then suggest that charismatic speakers are expected to present 'new' rather than 'old' information to listeners and are not expected to convey 'uncertainty' or 'incredulity' in their messages.

We also examined some larger intonational contours as possible correlates of charisma. In particular, we compared standard declarative contours ($H^*L-L\%$) with the most common of the downstepped contours in SAE ($H^*!H^*L-L\%$). Analyzing the distribution of these two contour types compared to all other contours in the stimuli, we found that the number of downstepped intermediate phrases in a token significantly and negatively influenced ratings of charisma ($r = -0.109$, $p = 0.0419$).[7] The more downstepped phrases, the less charismatic the token was rated. While the meaning conveyed by downstepped contours is an open research question, it has been associated in the literature with topic beginnings and endings, with the conveyance of 'given' information, and with the conveyance of already 'known' information in a markedly didactic way, and is often said to be used by teachers to imply that students should already know the information being presented (Dahan, 2002; Ladd, 1996; Pierrehumbert and Hirschberg, 1990). This contour has been anecdotally observed to offend some American listeners. Any of these interpretations presents plausible reasons for subjects to rate downstepped tokens low with respect to charisma.

We were also interested in investigating another of our hypotheses, that variation in intonational expressiveness – whether of pitch or of perceived loudness – might influence judgments of charisma. We had observed that effective speakers often use variation in these features to keep listeners engaged in what they are hearing. So we investigated the acoustic–prosodic features we had previously examined over *entire* tokens again, both at the level of the smaller intermediate phrase (ToBI level 3, 4 or 4 boundaries, inclusively) and at the intonational (Tobi level 4 only) level. We examined the number of such phrases in each token, the mean and standard deviation of the (normalized) maximum and mean pitch, and the mean and standard deviation of the intensity

[6] All downstepped pitch and phrase accents were merged with their non-downstepped versions.

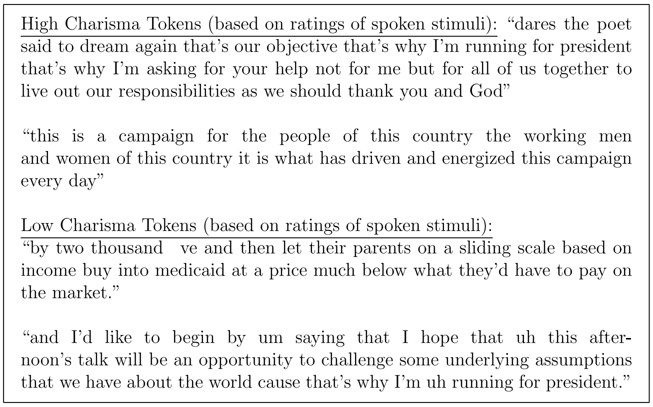
[7] The number of downstepped intermediate phrases was normalized by the total number of phrases in the token.

(calculated over segmentals only) across phrases within the token. For intermediate phrases, only the standard deviation of the normalized maximum pitch approached significant correlation with ratings of charisma ($r = 0.0781,\ p = 0.144$). That is, tokens whose individual intermediate phrases varied considerably in maximum pitch (i.e., in pitch range) were rated as more charismatic than those with less variation. For intonational phrases, the mean ($r = 0.128,\ p = 0.0166$) and standard deviation ($r = 0.111,\ p = 0.0361$) of the normalized maximum intensity as well as the number of words per phrase ($r = 0.111,\ p = 0.0358$) are all significantly and positively correlated with ratings of charisma. Such rapid change in pitch and intensity may well contribute to conveying such charisma-correlated attributes as passion and enthusiasm. The raw number of intermediate phrases within a token is also positively correlated with charisma ($r = 0.894,\ p = 0.0650$), while the mean number of intermediate phrases per intonational phrase approaches significance ($r = 0.0744,\ p = 0.164$). This is consonant with our findings that greater number of words and longer utterances were associated with higher ratings of charisma.

We note that, in many of the analyses presented above, we find significant linear correlations, either positive or negative, between acoustic–prosodic and lexico-syntactic features and ratings of charisma. However, while these linear relationships are confirmed by the data, the true nature of some of these interactions may be at least potentially U-shaped rather than truly linear. For example, we found mean pitch to correlate positively with reported perceptions of charisma. Naturally, there is a limit to this interaction. An unnaturally high pitch might be unlikely to be perceived as charismatic despite the evidence of correlation reported here. Identifying the limits of the linear correlation "sweet spot", the point at which a feature leads to a maximal perception of charisma, for these variables remains a topic for future study.

## 4. Charisma judgments from text

The results from the experiment on spoken data described in Section 3.3 indicated that there are both acoustic–prosodic and lexico-syntactic correlates to charismatic speech. In order to separate the influences of *what* is said from *how* it is said, we repeated the experiment described in Section 3 using transcripts of the speech tokens. To further investigate possible lexico-syntactic correlates of charisma in more detail, we also included some additional material generated originally in written form. A new group of subjects was recruited to rate transcriptions of the original spoken tokens, plus the additional text tokens, on the same set of 26 statements used in the first experiment. By collecting judgments on transcripts of the speech tokens used in the previous study, we hoped to distinguish the influence of the acoustic–prosodic features from lexical content and syntactic form. By using the same experimental design, we can determine if subjects employ a common definition of charisma (cf. Section 3.2.2) when judging text and when judging speech. In this section, we will describe the text-based experiment and compare the results of the speech study to the responses that were collected when subjects assessed corresponding transcripts.


High Charisma Tokens (based on ratings of spoken stimuli): "dares the poet said to dream again that's our objective that's why I'm running for president that's why I'm asking for your help not for me but for all of us together to live out our responsibilities as we should thank you and God"

"this is a campaign for the people of this country the working men and women of this country it is what has driven and energized this campaign every day"

Low Charisma Tokens (based on ratings of spoken stimuli):
"by two thousand five and then let their parents on a sliding scale based on income buy into medicaid at a price much below what they'd have to pay on the market."

"and I'd like to begin by um saying that I hope that uh this afternoon's talk will be an opportunity to challenge some underlying assumptions that we have about the world cause that's why I'm uh running for president."

Fig. 1. Sample transcripts of speech tokens.

### 4.1. Experimental design

To compare judgments of charisma from speech and from text, we orthographically transcribed the 45 spoken tokens used in the speech-based study (cf. Section 3).[8] Each of these tokens was approximately a single sentence long and was transcribed with all disfluencies explicitly indicated.[9] Mean number of words in the transcribed versions of the tokens was 28.73 words. Some examples appear in Fig. 1.

For further exploration of the role of lexico-syntactic cues in perceptions of charism, we also included a set of text materials selected from the original speakers' campaign speeches in addition to the transcribed spoken material. Thus, subjects in the text-based experiment rated more tokens (60 vs. 45) than subjects in the speech-based experiment. However, here we discuss only the speech transcript judgments. Presentation of these tokens was balanced as for the speech experiment.

Twenty three native American English speakers were presented with the text tokens in a web interface, as before. However, for this experiment, we asked subjects to come to Columbia Speech Lab and we compensated them for their participation; the sessions took place between December 14, 2004 and January 19, 2005. Again, using a web form similar to that shown in Appendix A, subjects were asked to rate their agreement with the 26 statements about the speaker of a given speech transcript on a five-point Likert scale, as described in Section 3.1. The only difference in the forms used in this text-based experiment was that the text segment to be rated was presented in the web browser,

[8] For consistency with the first study, we included the same duplicated token in this study.

[9] The median number of disfluencies was zero. The "most disfluent" token contained 12.1% disfluencies.

above the presentation of the statements to be rated. Again, the order of presentation of the 60 transcriptions was randomized for each subject and the order of the 26 statements was randomized for each token. There was no restriction placed on the order in which statements could be rated, although the subject could not continue on to judge the next token without rating all the statements with respect to the current token. At the end of the survey, subjects were again asked to indicate the names of any authors whose statements they thought they had recognized. It took subjects an average of ~2 h to complete the survey. The shortest time was 1 h 20 min; the longest, 2 h 50 min.

### 4.2. Text study analysis and comparison to speech findings

In this section, we compare subjects perceptions of charisma from speech transcripts of our spoken tokens, to their perceptions of charisma of the spoken tokens themselves, as discussed in Section 3, to see how modality affects charisma judgments. Our goals are to determine whether there are major differences in the perception of charisma which may be attributed to modality: is charisma more reliably conveyed in speech vs. text? Are speech-based features more influential in subject judgments of charisma than text-based features, or vice versa?

#### 4.2.1. Across-subject agreement on ratings

With regard to subjects' responses to the charismatic statement, the results from the speech tokens and from their transcripts are quite similar. Subjects showed an even lower degree of agreement with respect to the charismatic statement; agreement for this in text was $\kappa = 0.134$ and for speech $\kappa = 0.224$, although this was in line with the mean agreement over all statements (text: 0.148; speech: 0.213). It seems clear, then, that acoustic and prosodic information provides useful information for rating all statements.

#### 4.2.2. Within-subject correlation of statement ratings

Despite this low agreement, we observed that, in ratings of both spoken tokens and their transcripts, individual subjects' judgments of particular tokens were consistent with respect to the statements that were highly correlated with the charismatic statement. In both studies, when subjects "strongly agreed" with "The speaker is charismatic", they also agreed with the statements that "The speaker is enthusiastic", "...charming", "...persuasive", and "...convincing" (cf. Tables 3 and 4).

The text study subjects' 'functional' definition of a charismatic speaker shared four attributes with that of the speech-based judges: charm, enthusiasm, persuasiveness, and convincingness. "The speaker is charming" and "The speaker is persuasive" demonstrated the two highest correlations with "The speaker is charismatic" in both the speech and text experiments. However, where the speech raters also showed a positive correlation of charisma with passion and a negative correlation with boringness, the text-only group substituted positive correlations with believability and powerfulness in its 'top six'. Again, it seems likely to attribute these differences to the difficulty of conveying passion or boringness in text alone. In all, however, the agreement across modalities of this functional definition was striking. It is worth noting at this point that the most consistently correlated statements with the charismatic statement are consistent across all of the text tokens, both the transcribed speech tokens and the longer paragraph-lengthed tokens. There are slight differences in the exact kappa values, which lead to the cardinality of 'believable' and 'powerful' being inverted, but the set of the most consistent remains the same. This serves to reinforce the conclusion that this 'functional definition' is consistent not only regardless of modality of presentation, but also regardless of the stimuli being assessed.

Table 4
Statements with most consistent positive and negative correlation with charismatic statement based on ratings of transcribed speech tokens.

| Statement | $\kappa$ |
|---|---|
| The speaker is charming | 0.637 |
| The speaker is persuasive | 0.599 |
| The speaker is enthusiastic | 0.582 |
| The speaker is convincing | 0.574 |
| The speaker is believable | 0.560 |
| The speaker is powerful | 0.553 |

#### 4.2.3. Influence of speaker, topic, genre, and order of presentation on charisma ratings

Speaker-dependent characteristics are perforce limited in text compared to speech. Specifically, the lack of acoustic–prosodic information, forces subjects to evaluate the speaker on the merits of lexical qualities – syntax, word choice, semantics, pragmatics – alone. For the most part, those speakers who were rated as below average with respect to charisma based on the speech tokens were the same as those based on the corresponding transcripts. Mean scores for each speaker appear in Table 5. The similarities between the speaker ratings across presentation media suggest that lexical content is especially relevant to the communication of charisma. Otherwise, we would expect to observe greater differences between speaker perceptions across media.

For ratings from text alone the speaker of a transcript still significantly influenced judgments of charisma (ANOVA $p = 6.44 \times 10^{-13}$). While we had hypothesized in the speech experiment that some speakers would be perceived as more charismatic than others, it is interesting that such differences emerged even from text tokens, where there seems less opportunity for candidates to stamp material with their personal style, and there is a possibility of their texts to have multiple authors.

However, the identity of the most charismatic authors for raters from text differed slightly from the list for raters from speech: while Rep. Edwards, Rev. Sharpton and Gov. Dean ranked as the most charismatic speakers from speech, Sen. Kerry (mean rating 3.50), Gen. Clark (3.48), and Sen. Edwards (3.39) were ranked most highly by raters

Table 5
Mean ratings of charisma by speaker from speech vs. speech transcripts.

| Speaker | Transcript | Speech |
|---|---|---|
| Gen. Clark | 3.48 | 3.20 |
| Gov. Dean | 3.30 | 3.33 |
| Rep. Edwards | 3.39 | 3.75 |
| Rep. Gephardt | 3.05 | 2.77 |
| Sen. Kerry | 3.50 | 3.20 |
| Sen. Kucinich | 2.80 | 2.73 |
| Amb. Moseley Braun | 3.03 | 2.92 |
| Sen. Lieberman | 2.64 | 2.38 |
| Rev. Sharpton | 3.13 | 3.44 |
| Overall mean | 3.16 | 3.09 |

from text. And while Sen. Lieberman, Rep. Kucinich, and Rep. Gephardt ranked lowest from speech, Rep. Gephardt was replaced by Amb. Moseley Braun in text-based rankings, with ratings of: Sen. Lieberman (2.64), Rep. Kucinich (2.80), and Amb. Moseley Braun (3.102). Thus, two of three speakers who were rated highest and lowest for charisma from speech retained their positions when ratings were based on text. Rep. Edwards and Rev. Sharpton are the two speakers whose tokens are rated as significantly more charismatic in speech than in text. These two speakers are the only two who have obvious southern accents. This anecdotal observation suggests that there is some aspect of the southern accent that contributes to perceptions of charisma, potentially the acoustic–prosodic variations common in this regional accent or the segmental productions which characterize it.

Note also that exit interviews with subjects who rated from text showed much less perceived author recognition than post-survey questions of subjects who rated from speech. When raters were asked if they had identified any authors of the text tokens, the mean number of authors recognized was only 1.22 with a maximum of 4 and a minimum of 0. This compares with a mean of 3.25 from speech (Section 3.2.3), with a maximum of 6 and minimum of 0. In the text experiment, as in the speech experiment, subjects rated tokens from a recognized author as significantly more charismatic (mean rating 3.48) than those spoken by unrecognized individuals (mean rating 3.11). This difference is statistically significant ($t$-test $p = 7.49 \times 10^{-4}$). However, regardless of modality, when a subject believed they had recognized a speaker/author, they tended to rate him or her as more charismatic than unrecognized authors. These results demonstrate that subjects made decisions about their perceptions of charisma similarly, regardless of the medium through which the stimulus was presented.

As with the speech experiment, genre was a significant influence on subject perceptions of transcribed speech; tokens that were originally delivered as part of a stump speech, an interview, a debate or a campaign ad were rated quite differently with respect to charisma (ANOVA $p = 4.54 \times 10^{-13}$). As with the spoken originals, transcribed portions of stump speeches were consistently rated as more charismatic (mean = 3.34) than transcribed interviews (2.86), while ratings of debate transcripts (3.32) approximated the overall mean (3.16); the single transcribed portion of a campaign advertisement was rated as very charismatic (3.87), while its spoken counterpart had been rated in the previous experiment as very low in charisma compared to other genres (2.88). While this latter result is intriguing, we make no claims about the relationship between the campaign ad genre and charisma based on this singular token. As we discussed in Section 3.2.3, however, the observed relationship between genre and ratings of charisma for the other genres may be explained by the ease of expressing power, enthusiasm, and persuasion in a stump speech as opposed to an interview.

While in our perception study of speech tokens topic had no statistically significant impact on subject ratings of charisma, in the text-based study topic (postwar Iraq, health care, taxes, reason for running, and content-neutral) of the speech segment did have a statistically significant impact (ANOVA $p = 1.06 \times 10^{-9}$). Health care tokens obtained ratings of charisma (3.19) in line with the overall mean (3.16); tokens concerning President Bush's tax plan (2.97), and postwar Iraq (2.83) were rated lower than average; while content-neutral tokens (3.47) and candidates' reasons for running (3.35) were rated significantly above average.

#### 4.2.4. Lexico-syntactic properties of charismatic speech

To assess the role of lexico-syntatic features in transcripts compared with their influence when acoustic and prosodic information was also available, we examined the same lexico-syntactic features' correlation with charisma judgments when text tokens were presented as we explored for our speech tokens in Section 3.3.1. These were: the number of words in the token, the ratio of function to content words, the number of repeated words, Dowis' measure of lexical complexity, the token's pronoun'density', and the ratio of disfluencies to words in the token. We hypothesized that, if indeed lexico-syntactic factors play a role in others' perceptions of charisma, these features might exert even more influence over rater responses for the text survey vs. the speech study, since in the absence of acoustic and prosodic information these are the only factors available to the listener.

Over all, lexico-syntactic features which were highly correlated with charisma for the transcript study differed from those correlated with charisma in the speech study, both in direction of correlation and in degree. First, while we had found a tendency approaching significance ($r = 0.097$, $p = 0.068$) for number of words in token to increase subject ratings of charisma when spoken tokens were presented to subject, we found no such influence of number of words on subject ratings of charisma when the same tokens were presented in transcript form. While this difference might have been influenced to some extent by the modality itself – speech tokens were repeated continuously while subjects in the text study may well *not* have *read* the tokens an equal number of times. So, it may have been the amount of *exposure* to the stimulus rather than the length of the stimulus itself which led to the greater impact of number of words on charisma ratings – itself an intriguing possibility.

To address this possibility, we performed some further analysis on both the speech and the text studies. We calculated the number of times each token was played in the speech study before the subject had completed ratings of all statements, finding that, indeed, the more times a subject heard a token, the more charismatic they rated it ($r = 0.108$, $p = 0.046$). While we had no way of determining how many times a subject read a given transcript in the text study, we *did* find a significant positive correlation ($r = 0.127$, $p = 0.00564$) between the amount of time a subject spent responding to all of the statements for a given token with how charismatic they found the speaker to be. So, at least we can conclude in both studies that the more time a subject spent on a token, the more charismatic they rated the token, and this greater attention might possibly be due to increased attention to the token itself. However, further and controlled experimentation would be needed to verify this hypothesis.

We next examined the function word/content word ratio we had found to correlate with charisma ratings from speech (Section 3.3.1), this time for the text experiment. Again we found that, in contrast to previous hypotheses about the importance of the content of charismatic leaders' messages, tokens rated as more charismatic from transcripts also contained *fewer* content words (nouns, verb, adjectives, and adverbs) (for speech, $r = 0.102$, $p = 0.0569$ and for transcripts $r = 0.113$, $p = 2.27 \times 10^{-4}$). The greater the ratio of function to content words, the more charismatic the speaker was perceived to be in both modalities. In Section 3.3.1, we conjectured that this somewhat surprising result might also be related to the syntactic structures employed by a speaker: more charismatic speech may include more complex syntactic structures, perhaps to produce rhetorical effect. Such a difference may serve to explain the stronger correlation in the transcript study results than in the speech study. However, further study will be needed to test these conjectures.

Subject ratings of tokens in different modalities also differed in terms of the lexical complexity (mean syllables per word) measure proposed by Dowis (2000). While subjects who rated from transcript showed no influence of this feature on their ratings, those in the speech experiment did ($r = 0.123$, $p = 0.0210$). For spoken tokens, the greater the lexical complexity of the token, the more charismatic the speaker was considered. This relationship between spoken multi-syllable words and charisma is counter to claims in the literature that charismatic speech tends to be simple and straightforward. A possible explanation for this difference may be found in the cognitive psychology literature. Larson (2003) claims that listeners are forced – by the mode of transmission – to take more time to *hear* longer words in speech, while words read in text are read as a whole. Thus length – in syllables or letters – is not directly related to the amount of time needed for reading. If, as we conjectured previously, the length of time subjects attended to a message influenced their perception of its charisma, this may be another case of increased processing time leading subjects to rate tokens as more charismatic.

Comparing ratings of speech and transcript tokens with respect to the 'personal' nature of the tokens, we again found considerable differences between judgments from speech and from transcripts. While density of first person pronouns demonstrated a significant influence in both surveys, with speech stimuli ($r = 0.116$, $p = 0.0294$) and transcribed stimuli ($r = -0.0625$, $p = 0.0443$), this correlation was *positive* for speech tokens but *negative* for text tokens. That is, the use of first person pronouns appeared to cause subjects to rate tokens as *less* charismatic in text, but *more* charismatic in speech. This puzzling finding appears to contradict much of the literature on the sources of a leader's charisma. As noted above, two major components of charisma according to Boss (1976) (citing Weber (1947) and Davies (1954)) are "qualities residing in the ['leader–communicator'] himself" and "the perceived effect on the 'listener–followers'". The effect on the listener is the establishment of a relationship, not to the communicator's ideas so much as to the communicator him or herself. We thus were not surprised to find that personal speech manifested in the use of first person pronouns appeared to correlate with charisma judgments in the speech survey. However, it seems that subjects *reading* the same pronouns in text may have been 'put off' by this usage; perhaps readers expect a more formal style in written documents than in spoken data, especially from public figures.

Clarity of message, as measured by the absence of disfluency, was an important correlate of charisma judgments in our speech experiment, with the ratio of disfluencies to the total number of words in a token being *negatively* associated with ratings of charisma ($r = -0.124$, $p = 0.0204$). We found a similar correlation in judgments from text ($r = -0.148$, $p = 1.65 \times 10^{-6}$). While disfluencies decreased subjects' perception of tokens as charismatic in both experiments – perhaps indicating a lack of planning or uncertainty in speakers' message – the effect was greater when disfluencies were read than when they were heard. This effect may be a result of subjects' expectations of each of the genre. While speech normally contains disfluencies, which may not even be recognized as such (Bard and Lickley, 1996), text normally does not, and transcribed disfluencies thus become a clearly noticeable phenomenon.

Despite differences in some of the features that are correlated with charisma in the two modalities, we nonetheless note a large amount of similarity in other features. Thus it seems fair to conclude that both *what* is said and *how* it is said influence listener/readers' judgments of charisma. However, the different effects observed in lexical features between the two studies – speech tokens vs. transcripts – suggest something more complex than a simple additive model. If some lexical features, such as the use of first person pronouns and the 'complexity' of words differ dramatically in their relationship to charisma judgments between spoken and written versions of the same message, such differences suggest that spoken language and written language may be perceived quite differently with respect to their charismatic effect in certain important ways. Such differences may indicate broader differences in the perception of written and spoken genres, beyond our study of perceptions of charisma.

### *4.3. Text study analysis and results*

In this section, we present an analysis of the text study responses along the same lines as reported on the speech study in Section 3.2. Unless otherwise indicated, as in Section 4.2, the results reported here are based on subject responses on those text tokens that were constructed from transcribed speech tokens used in the speech study. Whenever possible, comparisons between these analyses and results and those from the speech study will be made.

#### *4.3.1. Across-subject agreement on ratings*

We again used the weighted kappa statistic (Cohen, 1968) with quadratic weighting to determine the inter-subject agreement with respect to their assessments of each statement on each transcript. The mean pairwise $\kappa$ value over all 45 segments and 26 statements was 0.148, somewhat lower than for the speech tokens discussed in Section 3.2.1 ($\kappa = 0.218$). This agreement is quite low. As in our analyses of the speech experiment responses (Section 3.2.1), we analyzed the kappa contribution of each token in order to identify any sources of variation. The differences in kappa across all tokens was not significant. However, by examining the kappa contribution from each of the 26 statements individually, we could determine which of the statements are the most and least consistently ranked across subjects. As in the case of the speech stimuli, there was a substantial amount of inter-annotator agreement on the text stimuli with respect to the individual statements as well, although far less than for the speech stimuli. Tables 6 and 7 contain the five statements with the highest and lowest kappa scores.

Table 6
Most consistent statements with respect to inter-subject agreement in text survey.

| Statement | $\kappa$ |
|---|---|
| The speaker is accusatory | 0.280 |
| The speaker is angry | 0.263 |
| The speaker is friendly | 0.197 |
| The speaker is knowledgeable | 0.193 |
| The speaker is confident | 0.179 |

Table 7
Least consistent statements with respect to inter-subject agreement in text survey.

| Statement | $\kappa$ |
|---|---|
| The speaker is spontaneous | 0.0388 |
| The speaker is desperate | 0.0481 |
| The speaker is boring | 0.0508 |
| The speaker is ordinary | 0.0640 |
| The speaker is threatening | 0.0939 |

The two most consistently rated statements, “The speaker is accusatory” and “The speaker is angry”, were the most basic statements of the 26 to which subjects had to respond. Anger is a basic emotion and accusation is a common speech act. While their perception of the anger of a speaker, or how accusatory a statement is, may vary, we can assume they all meant the same thing when they used the terms “angry” and “accusatory”. This cannot be said, for example, about “The speaker is ordinary”; what is ordinary to one person may be extraordinary to another, regardless of the speech presented, the subjects may understand the statement in fundamentally different ways. That being said, significant disagreement regarding perceptions of these two concepts persisted – kappa of 0.280 represents fairly low agreement.

The kappa scores of the bottom five statements represented agreement not much greater than chance. These results can be attributed to one of two sources. Either there was nothing in the stimuli that systematically influenced subject perceptions, or subject responses to the stimuli show strong individual differences. In the cases of “The speaker is desperate” and “The speaker is spontaneous”, it is reasonable to assume that a transcript might not be able to reliably transmit this information about a speaker. On the other hand, assessments of how “boring” or “ordinary” a speaker is may be very subjective.

When we compare these findings to ratings of our spoken stimuli (Table 1), we see that, for both sets of stimuli, “The speaker is accusatory” was the statement on which there is strongest inter-rater agreement; kappa was 0.512 for the speech survey and 0.280 for the text experiment. However, it appears that subjects more easily agreed when rating tokens from speech than from text: agreement on the top five statements for the speech experiment ranged from 0.362 to 0.512, while agreement for the text study ranged from only 0.179 to 0.280 for the top five most agreed upon statements. We may conjecture that these differences indeed have arisen from the difference in the modality of the tokens being rated. For example, it may have been particularly difficult to rate text tokens in terms of spontaneity, boringness or enthusiasm with only text available.

Note also that, while the charismatic statement was the eighth most consistently rated statement from speech, with a kappa of 0.224 (Section 3.2.1), in the text experiment the charismatic statement had $\kappa = 0.132$. This placed it as only the nineteenth most consistently labeled statement. So, we may conjecture that charisma is more easily agreed upon from speech than from text alone.

#### *4.3.2. Influence of topic and genre on charisma*

Whether transcribed speech was originally delivered as part of a stump speech, an interview, a debate or a campaign ad significantly influences subject ratings of charisma (ANOVA $p = 4.54 \times 10^{-13}$). Transcribed portions of stump speeches were consistently rated as more charismatic (mean = 3.34) than transcribed interviews (2.86), while ratings of debate transcripts (3.32) approximated the overall mean (3.16). The single transcribed portion of a campaign advertisement was rated as very charismatic (3.87); while this latter result is intriguing, we make no claims about the relationship between the campaign ad genre and charisma based on this singular token. As we discussed in Section 3.2.3, however, the observed relationship between genre and ratings of charisma for the other genres may be explained by the ease of expressing power, enthusiasm, and persuasion in a stump speech as opposed to an interview.

The topic (postwar Iraq, health care, taxes, reason for running, and content-neutral) of the speech segment had a statistically significant impact (ANOVA $p = 1.06 \times 10^{-9}$) on subject ratings of charisma from text. Health care tokens demonstrated ratings of charisma (3.19) in line with the overall mean (3.16). Tokens concerning President Bush's tax plan (2.97), and postwar Iraq (2.83) were rated lower than average, while content-neutral tokens (3.47) and candidates' reasons for running (3.35) were significantly above average. These results support the claim in the sociological literature that a charismatic leader and his or her rhetoric must have a special relevance to reality, in order to command attention and followers; Boss (1976) finds this to occur when a leader possessing a "shared history" with his followers arises with a "mission-crusade" to successfully resolve a "crisis". In other words, charisma does not exist in a vacuum; what may have been charismatic in the past may be ineffective in the future. The stimuli were all generated before or during the Democratic Primary season, late 2003 to early 2004. However, the text survey was administered between 12/14/04 and 1/19/05, approximately 1 year later. During this year public sentiment about Iraq had significantly changed, and the relevance of discussing President Bush's by then established tax plan had probably diminished significantly. This difference in real-world context could easily have rendered the transcripts concerning these topics less likely to be deemed charismatic.

### 4.3.3. *Influence of order of presentation on charisma*

Recall that, in the preparation of stimuli for the speech survey, an error resulted in a speech token being presented twice (cf. Section 3.1). In order to compare results from the speech and transcript survey reliably, this double presentation was retained for the text survey. We are, therefore, again able to measure subject consistency by comparing ratings on two different presentations of the same token. Here, as in the speech study, we found no significant variation in subject ratings between the two presentations, suggesting at least that subject judgments were reliable and that there were no priming effects. With only a single data point of course, this finding requires further confirmation in later studies.

## 5. Conclusions

The experiments described in this paper represent a first step toward an empirically based understanding of the communication of charisma through speech and text. Our comparison of subject ratings of charisma and other personal attributes in response to tokens of spoken and transcribed political statements confirms some claims found in the previous literature but challenges others.

Our studies show that there are substantial differences in subject perceptions of charisma. In particular, there was no strong consensus among our raters as to what stimuli are perceived as charismatic and which are not. However, subject responses to the stimuli demonstrated a consistent, within-subject, functional definition of charisma, in that tokens they judged charismatic were also judged similarly on other dimensions. This definition was employed similarly by subjects across experiments and indicates that, for most subjects, charismatic speakers are also perceived as enthusiastic, charming, persuasive, and convincing.

In general, tokens that were heard to be charismatic, were also judged charismatic when read. This finding suggests a substantial influence of the lexical, syntactic, semantic and/or pragmatic content of speech on the communication of charisma. However, the fact that some of the same lexical characteristics that proved strong predictors of charisma in speech showed quite different effects in our written survey indicates that there are also important differences in how charisma – and perhaps other speaker qualities – may be perceived in the two modalities. For example, the length of speech tokens was positively correlated with charisma judgments in speech, but not in text, suggesting a basic difference perhaps related to the amount of exposure to subjects of the speaker's message. Also, the use of personal pronouns was positively correlated with charisma judgments from speech, but negatively correlated from text, suggesting that certain types of language may be deemed more appropriate in different modalities. We further observed a number of acoustic–prosodic properties that were also highly correlated with charisma, indicating that *how* the speech is spoken plays an important role in the communication of charisma. These properties included a faster speaking rate, speech that occurred higher in the pitch range, and varied with respect to pitch and amplitude – all aspects of speech commonly associated with a more engaged and lively style of speech and all predicting higher ratings of charisma.

We also found that, regardless of the medium of stimulus, recognized speakers were perceived as more charismatic those not recognized. In the analyses presented in this paper, we examined speech material exclusively, whether spoken or transcribed. This result highlighted a difficult aspect of studying charisma. Determinations of charisma are performed within a larger context than simply the speech token itself. Knowing who the speaker is can very likely influence judgments, even though we found no correlation of charisma judgments with subjects' agreement with what the speaker of a token had said. Additionally, we found evidence suggesting that, when the relevance of topics to current events changes over time, statements about them may change radically in their capacity to influence subjects' perception of speakers' charisma. More empirical work is necessary to control for both these factors, to help us understand the role that context as well as lexico-syntactic and acoustic–prosodic features play in perceptions of charisma.

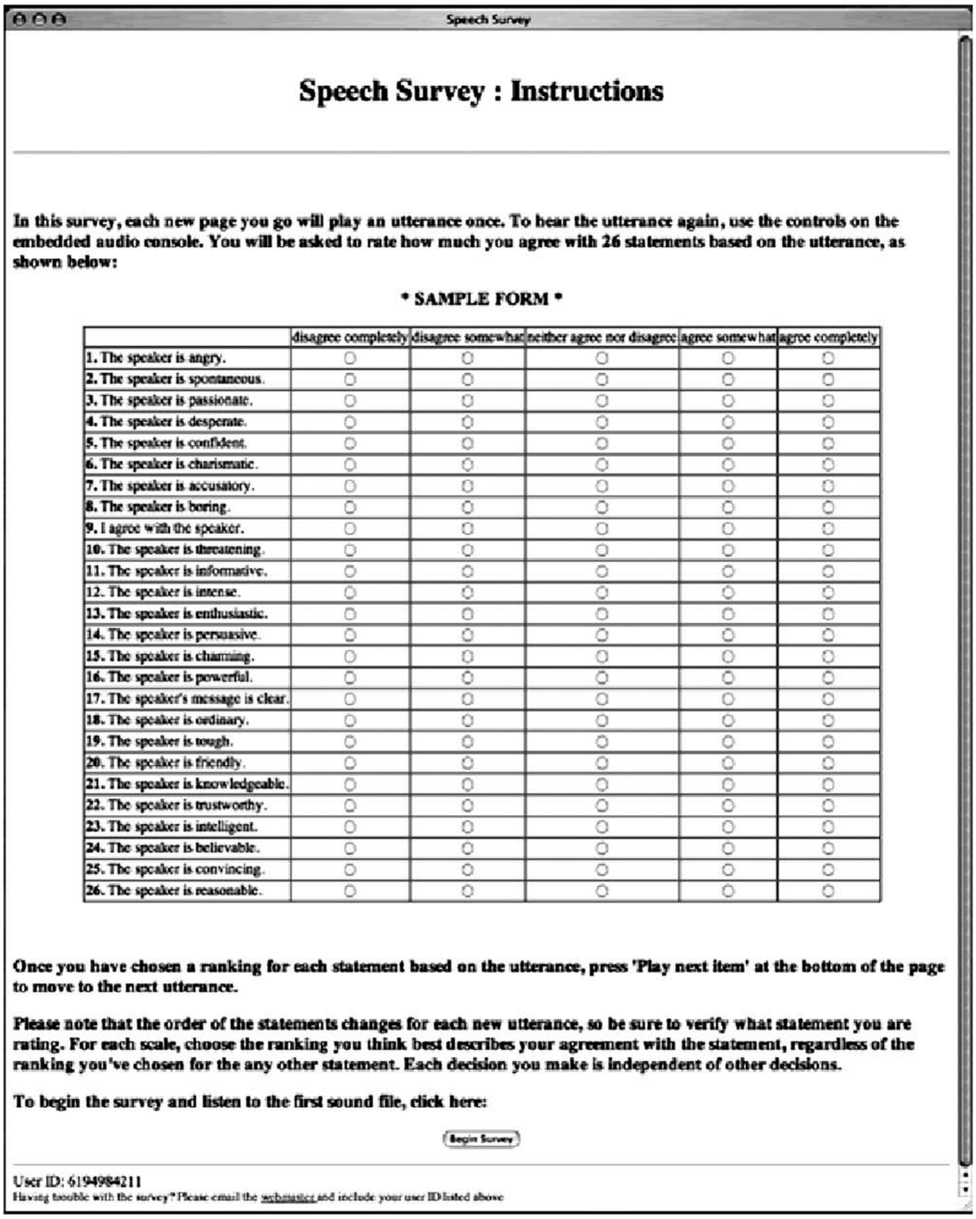
## 6. Future research

While our study of U.S. political speech has enabled us to discover some important potential correlates of charisma in the language speakers use, it is clear that charismatic speech is a phenomenon deeply steeped in the individual culture of speakers and hearers. To examine differences and similarities across cultures, we are currently collecting similar charisma judgments from native speakers of Palestinian Arabic. Our stimuli are drawn from Palestinian Arabic political speech. Our goals are to study how charisma is communicated in Palestinian Arabic, and to identify the similarities and differences between American English and Palestinian Arabic communication of charisma. We are also testing the judgments of non-native speakers on both our English and Arabic spoken stimuli, to compare subjects' charisma judgments of speech in their own language with speech in an unfamiliar language. Do native speakers of SAE, for example, use acoustic–prosodic features to make charisma judgments in Arabic that are similar to those they make in English? This research may help us to understand what is general about human perceptions of charisma and what is more culturally determined.

## Acknowledgements

The authors would like to thank Fadi Biadsy, Frank Enos, Jackson Liscombe, Judd Sheinholtz, Aron Wahl, and Svetlana Stenchikova for their help in preparation of the experiments described in this paper. This work was supported by a grant NSF-IIS-0325399 from NSF.

## Appendix A. Sample web form for speech survey



Speech Survey

**Speech Survey : Instructions**

**In this survey, each new page you go will play an utterance once. To hear the utterance again, use the controls on the embedded audio console. You will be asked to rate how much you agree with 26 statements based on the utterance, as shown below:**

*** SAMPLE FORM ***

| | disagree completely | disagree somewhat | neither agree nor disagree | agree somewhat | agree completely |
|---|---|---|---|---|---|
| 1. The speaker is angry. | ○ | ○ | ○ | ○ | ○ |
| 2. The speaker is spontaneous. | ○ | ○ | ○ | ○ | ○ |
| 3. The speaker is passionate. | ○ | ○ | ○ | ○ | ○ |
| 4. The speaker is desperate. | ○ | ○ | ○ | ○ | ○ |
| 5. The speaker is confident. | ○ | ○ | ○ | ○ | ○ |
| 6. The speaker is charismatic. | ○ | ○ | ○ | ○ | ○ |
| 7. The speaker is accusatory. | ○ | ○ | ○ | ○ | ○ |
| 8. The speaker is boring. | ○ | ○ | ○ | ○ | ○ |
| 9. I agree with the speaker. | ○ | ○ | ○ | ○ | ○ |
| 10. The speaker is threatening. | ○ | ○ | ○ | ○ | ○ |
| 11. The speaker is informative. | ○ | ○ | ○ | ○ | ○ |
| 12. The speaker is intense. | ○ | ○ | ○ | ○ | ○ |
| 13. The speaker is enthusiastic. | ○ | ○ | ○ | ○ | ○ |
| 14. The speaker is persuasive. | ○ | ○ | ○ | ○ | ○ |
| 15. The speaker is charming. | ○ | ○ | ○ | ○ | ○ |
| 16. The speaker is powerful. | ○ | ○ | ○ | ○ | ○ |
| 17. The speaker's message is clear. | ○ | ○ | ○ | ○ | ○ |
| 18. The speaker is ordinary. | ○ | ○ | ○ | ○ | ○ |
| 19. The speaker is tough. | ○ | ○ | ○ | ○ | ○ |
| 20. The speaker is friendly. | ○ | ○ | ○ | ○ | ○ |
| 21. The speaker is knowledgeable. | ○ | ○ | ○ | ○ | ○ |
| 22. The speaker is trustworthy. | ○ | ○ | ○ | ○ | ○ |
| 23. The speaker is intelligent. | ○ | ○ | ○ | ○ | ○ |
| 24. The speaker is believable. | ○ | ○ | ○ | ○ | ○ |
| 25. The speaker is convincing. | ○ | ○ | ○ | ○ | ○ |
| 26. The speaker is reasonable. | ○ | ○ | ○ | ○ | ○ |

**Once you have chosen a ranking for each statement based on the utterance, press 'Play next item' at the bottom of the page to move to the next utterance.**

**Please note that the order of the statements changes for each new utterance, so be sure to verify what statement you are rating. For each scale, choose the ranking you think best describes your agreement with the statement, regardless of the ranking you've chosen for the any other statement. Each decision you make is independent of other decisions.**

**To begin the survey and listen to the first sound file, click here:**

User ID: 6194984211

Having trouble with the survey? Please email the webmaster and include your user ID listed above

## References

Bard, E.G., Lickley, R.J., 1996. On not recognizing disfluencies in dialog. In: Proceedings of the ICSLP 96, pp. 1876–1879.

Bird, F.B., 1993. Charisma and leadership in new religious movements. In: Bromley, D.G., Hadden, J.K. (Eds.), Handbook of Cults and Sects in America, Vol. B, pp. 75–92.

Boersma, P., 2001. PRAAT, a system for doing phonetics by computer. Glot Internat. 5 (9/10), 341–345.

Boss, P., 1976. Essential attributes of charisma. South. Speech Comm. J. 41 (3), 300–313.

Brill, E., 1995. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Comput. Linguist. 21 (4), 543–565.

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220.

Cohen, R., 1987. Analyzing the structure of argumentative discourse. Comput. Linguist. 13 (1–2), 11–24.

Dahan, D. et al., 2002. Accent and reference resolution in spoken-language comprehension. J. Memory Lang. 47, 292–314.

Davies, J., 1954. Charisma in the 1952 campaign. Am. Polit. Sci. Rev. 48.

Dowis, R., 2000. The Lost Art of the Great Speech. AMACOM, New York.

Hamilton, A., Stewart, B., 1993. Extending an information processing model of language intensity effects. Comm. Quart. 41 (2), 231–246.

Ladd, D.R., 1996. Intonational Phonology. Cambridge University Press, Cambridge.

Larson, K., 2003. The science of word recognition or how I learned to stop worrying and love the bouma. <http://www.microsoft.com/typography/ctfonts/WordRecognition.aspx> (last accessed 09.01.05).

Marcus, J., 1967. Transcendence and charisma. West. Polit. Quart. 14, 237–241.

Pierrehumbert, J., Hirschberg, J., 1990. The meaning of intonational contours in the interpretation of discourse. In: Cohen, Morgan, Pollack (Eds.), Intentions in Communication. MIT Press.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. ToBI: a standard for labeling english prosody. In: Proceedings of the ICSLP'92, vol. 2, pp. 867–870.

Touati, P., 1993. Prosodic aspects of political rhetoric. In: ESCA Workshop on Prosody, pp. 168–171.

Tuppen, C., 1974. Dimensions of communicator credibility: an oblique solution. Speech Monogr. 41 (3), 253–260.

Weber, M., 1947. The Theory of Social and Economic Organization.