A Cross-Cultural Comparison of American, Palestinian, and Swedish Perception of Charismatic Speech

Fadi Biadsy, Andrew Rosenberg*, Rolf Carlson †, Julia Hirschberg*, and Eva Strangert ‡*

*Department of Computer Science Columbia University, New York, USA
{fadi; amaxwell; julia}@columbia.edu
†CSC, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
rolf@speech.kth.se
‡Department of Comparative Literature and Scandinavian Languages, Umeå University, Sweden
eva.strangert@nord.umu.se

Abstract

Perception of **charisma**, the ability to influence others by virtue of one's personal qualities, appears to be influenced to some extent by cultural factors. We compare results of five studies of charisma speech in which American, Palestinian, and Swedish subjects rated Standard American English political speech and Americans and Palestinians rated Palestinian Arabic speech. We identify acoustic-prosodic and lexical features correlated with charisma ratings of both languages for native and non-native speakers and find that 1) some acoustic-prosodic features correlated with charisma ratings appear similar across all five experiments; 2) other acoustic-prosodic and lexical features correlated with charisma appear specific to the language rated, whatever the native language of the rater; and 3) still other acoustic-prosodic cues appear specific to both rater native language and to language rated. We also find that, while the absolute ratings non-native raters assign tend to be lower than those of native speakers, the ratings themselves are strongly correlated.

1. Introduction

According to Weber [1], charismatic leaders are those who owe their power to their personal qualities and not to formal political or military institutions. One observation often made of charismatic leaders is their remarkable communicative skill. In previous studies, we have investigated the role that spoken language plays in subject perceptions of charismatic speech when native speakers rate speakers of Standard American English (SAE) and Palestinian Arabic speech [3, 8] and how these perceptions differ between the two cultures. In this study we investigate the acoustic-prosodic and lexical correlates of charisma ratings when speakers from one culture assess speech from speakers of another. We compare results of our earlier studies with perception studies of Palestinian and Swedish native speakers judging SAE speech and American speakers judging Arabic, to see how these correlates compare across cultures. What features do subjects appear to rely upon in judging charisma in a second language? Are these features similar to those they use when judging their own language?

In Section 2 we describe previous work on charisma and charismatic speech. In Section 3, we describe our materials and experimental design. We analyze subject judgments in Section 4. Acoustic/prosodic and lexical correlates to subject charisma ratings are presented in Section 5 and further cross-cultural comparisons in Section 6. We conclude in Section 7 and present directions for future work.

2. Previous Work

The qualities that charismatic leaders exhibit have been studied by researchers in rhetoric and social science [1]. Although these qualities are difficult to define, several authors have proposed attributes which charismatic individuals typically exhibit [2, 4]. Recently, we [3, 8] have investigated specific acoustic, prosodic, and lexical characteristics of charismatic speech, identified through a series of perception experiments designed to elicit native speakers' judgments of charisma in SAE and Palestinian Arabic speech. We found that the most consistent attributes correlated with charisma according to the American subjects were *persuasive*, *charming*, *passionate*, *convincing*, and **neither** *boring* **nor** *ordinary*, while Palestinian subjects viewed speakers who are *tough*, *powerful*, *persuasive*, *charming*, *enthusiastic*, and **neither** *boring* **nor** *desperate* to be most charismatic. In terms of acoustic-prosodic and lexical features, we found, inter alia, that in both cultures longer amounts of speech, and generally, dynamic speech produced with significant change in speaking rate, high in the speaker's pitch range, and with variation in intensity across intonational phrases was perceived as charismatic. Also, the use of simple sentences with repeated words increased charisma ratings, while disfluencies inhibited them. Below we describe the materials and experimental design of these previous experiments and three additional cross-cultural experiments using the same materials and design, in which **non**-native speakers rated SAE and Arabic.

3. Materials and Experimental Design

We chose materials from the 9 candidates (1 F, 8 M) seeking the Democratic nomination for U.S. president in 2004, to confine the range of opinions presented in the tokens, as the literature suggests that a listener's agreement with a speaker may affect charisma judgments [1, 2, 4]. To control for effect of topic, we included five tokens from each speaker, one each on: healthcare, postwar Iraq, President Bush's tax plan, the candidate's reason for running, and a neutral greeting. Since the tokens were recorded under a variety conditions, we normalized them for intensity to -12dBFS. We balanced tokens that "sounded charismatic" for each speaker with those that did not, relying upon judgments of four native SAE speakers. We selected a total of 45 speech segments of 2–28s duration, with a mean of

10s.

The Palestinian Arabic materials consist of 44 speech tokens of 3–28s duration, with a mean of 14s, two from each of 22 male native Palestinian speakers recorded from television programs on the Al-Jazeera News Channel web site (http://www.aljazeera.net) in 2005. Speakers and topics were varied as in the first corpus. Topics included the assassination of the Hamas leader, the debate among the Palestinian groups, the Intifada and resistance, the Israeli separation wall, the Palestinian Authority, and calls for reforms. We chose one token from each speaker that "sounded charismatic" to 3 native Palestinian informants and one that did not.

In the first two experiments, 12 (6 F, 6 M) American and 12 (6 F, 6 M) Palestinian subjects were presented with speech tokens in their native language [3, 8]. The experiments differed only in the materials presented to subjects for judgment — SAE or Arabic. Subjects were asked to rate each speaker on 26 statements using a five-point Likert scale in a webform survey. In addition to rating the charisma of each token's speaker (*the speaker is charismatic*), subjects were asked to rate speakers on a number of other attributes that have been associated with charisma in the literature, e.g. *the speaker is angry*.[1] Tokens were presented simultaneously with the the statements and repeated with 2 seconds of silence between iterations until the subject had responded to all 26 statements and moved to the next token. Order of presentation of tokens was randomized for each subject and the order of the 26 statements was randomized for each token. At the end of the survey, subjects were asked to list the names of any speakers they thought they had recognized.

For the current research three additional studies were conducted to examine how perceptions of charisma differ when subjects are presented with stimuli spoken in a language other than their native language. Using the experimental paradign described above, we asked 9 (6 F, 3 M) English-speaking native Swedish speakers to perform the SAE-based experiment. In two shorter experiments, we asked 12 (3 F, 9 M) English-literate native Palestinian Arabic speakers to rate only the charisma of the SAE tokens and 12 (3 M, 9 F) non-Arabic-literate SAE speakers to judge the charisma of the Arabic tokens.

4. Analysis of Subject Judgments

For each study, we examine subject agreement on ratings for all tokens, including the charismatic statement.[2] Ratings of charisma by American subjects on SAE show a mean κ of 0.232. Interestingly, these subjects demonstrate higher agreement when rating Arabic (mean κ: 0.383), suggesting that lexical, syntactic and semantic cues available to Americans when rating SAE stimuli may be a source of **dis**agreement in charisma judgments. However, Palestinian subjects demonstrate higher agreement (κ=0.348) when judging the charisma of speech in **their** native language than when rating SAE (κ=0.185). Agreement among Swedish subjects on the charisma of SAE is κ=0.226. The rather low level of agreement for all five studies may be due to the fact that the task conflates two factors: subjects' understanding of the concept of 'charisma' and subjects' identification of that concept in the speech of individuals. Agreement among subjects rating foreign speech may also be influenced by differences in their exposure and experience with the language.

[1] Cf. [3, 8] for the full list of statements.

[2] The weighted kappa statistic [5] with quadratic weighting was used to determine inter-subject agreement.

In all five experiments, we find that the speaker of a segment significantly influences subjects' ratings of charisma.[3] Subjects reports of recognized speakers, however, vary significantly over the experiments. Americans rating SAE tokens report recognizing 5.8 of 9 speakers on average and rate tokens spoken by a (purportedly) recognized speaker as significantly more charismatic (mean rating 3.39) than those spoken by unrecognized speakers (3.0). This may imply that familiarity with a speaker positively influences perceptions of charisma, or that charismatic speakers are more recognizable than uncharismatic speakers. However, in the other four studies, the identification rate of the speakers by our subjects is extremely low: the mean number of Arabic speakers reportedly recognized by Palestinian subjects is 0.55 of 22. No American subject identified any Palestinian speaker. The mean number of American speakers reportedly identified by Swedish subjects is 0.11 and by Palestinian subjects, 0.33. Unfortunately, there is not enough data here to draw conclusions about the influence of speaker recognition on charisma in non-native speech, but this will be an object of future research.

The topic of the tokens also has an effect on subjects' ratings of charisma. We see an effect approaching statistical significance (p=.052) on charisma ratings when Americans rate SAE stimuli, and a statistically significance effect in each of the other four studies: Americans rating Arabic stimuli (p=.0052), Palestinians rating Arabic (p=.043), Palestinians rating English (p=.0079); and Swedish speakers rating English (p=.0001). This may imply that the sensitivity and importance of a topic may influence either the emotional state of the speaker or of the rater.

5. Feature Analysis

We next extract acoustic-prosodic and lexical features from the experiment stimuli and, using linear regression, seek to identify characteristics of the stimuli that significantly correlate with subject ratings of charisma. Our stimuli are also ToBI labeled to see how categorical representations of prosody correlate with charisma judgments.[4]

5.1. Acoustic-Prosodic Features

A number of acoustic features show significant correlation with charisma in all five experiments. Mean pitch (r_e=.24; r_{pe}=.13; r_{aa}=.39; r_a=.2; r_s=.2),[5] mean (r_e=.21; r_{pe}=.14; r_{aa}=.35; r_a=.21; r_s=.18) and standard deviation (r_e=.21; r_{pe}=.14; r_{aa}=.34; r_a=.19; r_s=.18) of rms intensity over intonational phrases, and token duration (r_e=.09; r_{pe}=.15; r_{aa}=.24; r_a=.30; r_s=.12) all positively correlate with charisma ratings, regardless of the subject's native tongue or the language rated. Pitch range[6] is positively correlated with charisma in all experiments (r_e=.2;

[3] All p-values in Section 4 were determined by one-way ANOVA with repeated measures and are significant at the $p < .001$ level unless otherwise noted.

[4] While there is no current ToBI standard for Palestinian Arabic, we are developing one and use our current draft for this annotation. We also use HiF0 values as an alternate method of calculating speakers' pitch ranges over an intermediate phrase and have extract pitch and intensity dynamics across ToBI intermediate phrases using hand-annotated phrase boundaries.

[5] r_e refers to the correlation coefficient for the SAE experiment, r_a for the Arabic, r_{pe} for the Palestinians rating SAE, r_{aa} for Americans judging Arabic, and r_s for the Swedish study. Throughout, p-values are significant at the .05 level or better, except as otherwise noted.

[6] Calculated as the mean HiF0 of intermediate phrase. In ToBI, HiF0 indicates the location of the highest accented pitch peak within an intermediate phrase.

r_{pe}=.12; r_{aa}=.36; r_a=.23; r_s=.19). Looking at the role of intonational contour in charisma, again from our ToBI annotations, we see that, over all experiments, the proportion of words accented with a downstepped pitch accent (!H*) is positively correlated with charisma (r_e=.19; r_{pe}=.17; r_{aa}=.15; r_a=.25; r_s=.14), while the proportion of low pitch accents (L*) is significantly negatively correlated (r_e=-.13; r_{pe}=-.11; r_{aa}=-.25; r_a=-.24) — for all but Swedish judgments of SAE (r=-.04; p=.4). Downstepped contours are often associated with public or professorial speech in SAE, while L* accents often mark yes-no questions, which may explain these correlations for American raters. However, further research will be needed to understand the Swedish and Palestinian results. Generally though, across cultures, charisma judgments tend to correlate with higher f0, higher and more varied intensity, longer duration of stimuli, and downstepped (!H*) contours. The presence of disfluency (filled pauses and self-repairs) on the other hand, is **negatively** correlated with charisma judgments in all cases (r_e=-.18; r_{pe}=-.22; r_{aa}=-.39; r_a=-.48), except for Swedish judgments of SAE, where there is only a tendency (r=-.09; p=.087).

Other features which are correlated with charisma in some experiments but not in others indicate that, in general, all three language groups rating SAE pattern similarly — as do both groups rating Arabic. In fact, subjects agree upon language-specific acoustic-prosodic indicators of charisma, despite the fact that these indicators differ in important respects from those in the raters' native language. American subjects' similarity to Palestinian ratings is particularly striking, since, while Swedish and Palestinian subjects had some knowledge of SAE, the American subjects had no knowledge of Arabic. For example, for all groups rating SAE, minimum f0 — possibly indicating speech spoken in a higher overall range — is **positively** correlated with charisma (r_e=.14; r_{pe}=.15; r_s=.21), while Palestinian charisma judgments of Arabic **negatively** correlate with this features (r=-.16), and there is **no correlation** for Americans judging Arabic. Both groups judging Arabic rate speech more charismatic that exhibits larger standard deviations in f0 (r_a=.22; r_{aa}=.20) — varying more widely — but none of the groups judging SAE show the same effect. The presence of simple H* pitch accents, the most common pitch accent in SAE and used in standard 'declarative' contours, is **negatively** correlated with charisma for all SAE experiments (r_e=-.11; r_{pe}=-.14; r_s=-.12), but not for Arabic. Both groups rating Arabic show **positive** correlations with charisma for maximum intensity (r_a=.24; r_{aa}=.21) and standard deviation of intensity (r_a=.16; r_{aa}=.17), i.e. louder utterances but with more variation in loudness, but, of those who rated SAE, only Swedish subjects show any correlation between these intensity features and charisma ratings, and that correlation is **negative** for both (r_s=-.14; r_s=-.12).

Still other correlations of acoustic-prosodic features with charisma ratings do appear particular not only to the native language of rater but also to the language rated. For example, speaking rate[7] is **positively** correlated with charisma judgments only for American and Swedish ratings of SAE (r_e=.17; r_s=.16): the faster the speech, the more charismatic the speaker. However, when Palestinians judge Arabic speakers, rate approaches a **negative** correlation with charisma (r=-.08, p=.08), with **no** correlation between rate and charisma when Palestinians judge SAE or Americans judge Arabic. So, while some acoustic-prosodic features appear important to charisma decisions across languages, and others appear to be important language-specific cues — even to non-native speakers — yet other features appear to depend both on the raters' native language **and** the language rated.

[7]Calculated as the mean of the ratio of voiced to unvoiced frames.

5.2. Lexical Features

In earlier studies we reported a number of lexical correlates of charisma judgments based on American and Palestinian judgments of native speech [3, 8]. As with acoustic-prosodic features, we find again that at least some non-native judgments of SAE and Arabic resemble native judgments. For American and Swedish judges of SAE, the number of third person plural pronouns in a token, a possible 'distancing' device, is negatively correlated with charisma (r_e=-.19; r_s=-.16), while for all raters of SAE, the presence of inclusive first person plural pronouns (r_e=.16; r_{pe}=.13; r_s=.14), third person singular pronouns (r_e=.16; r_{pe}=.17; r_s=.15), and the percentage of repeated words — a rhetorical device conveying emphasis — (r_e=.12; r_{pe}=.16; r_a=.22; r_s=.18) is positively correlated with charisma, while the ratio of adjectives to all words is negatively correlated (r_e=-.12; r_{pe}=-.25; r_s=-.17). For judgments of Arabic, both Americans and Palestinians judge tokens with more third person plural pronouns (r_{aa}=.29; r_a=.21) and nouns in general (r_{aa}=.09; r_a=.1) as **more** charismatic.[8] This similarity of native and non-native raters in (some) lexical correlations with charisma is particularly puzzling for American ratings of Arabic, since, as noted above, no American raters were familiar with that language. We hypothesize that these lexical features may themselves be correlated with acoustic-prosodic features, which should be more available to non-native judges.

6. Charisma Ratings Across Cultures

To examine perceptions of charisma across cultures more generally, we compare charisma judgments between each pair of groups rating the same stimuli (e.g., American with Arabic speakers rating SAE material). For each group of raters, we construct a single charisma rating for each token: the mean of individual subject responses to the charisma statement. Using these pairs of aggregated scores, we perform a paired t-test to compare the ratings of the same tokens by the two groups. Next, to see whether any differences we may observe can be attributed to subjects' different use of the rating scale, we normalize the charisma scores by raters (using z-score normalization) and calculate the correlation between the aggregated normalized scores for each pair of rater groups.

The means of the American and Palestinian ratings of SAE tokens are 3.19 and 3.03, respectively; this difference is not significant. However the correlation of z-score-normalized charisma ratings **is** significant and positive (r=.47). These groups' ratings of SAE thus are **not** significantly different in raw scores and **are** correlated with one another. For them, there appears to be no notable difference in use of the Likert scale and a significant agreement over all in ratings for each token. Similarly, we find no significant difference between the ratings of Swedish (mean: 3.01) and Palestinian (mean: 3.03) subjects rating SAE and again the correlation between the groups is significant (r=.55), indicating that both groups are ranking the tokens similarly with respect to charisma.

Comparing American with Swedish judgments of SAE, we **do** find a significant difference in means (American mean: 3.19; Swedish: 3.01), but again the aggregate normalized scores from these subject groups **do** correlated significantly and positively

[8]Note that Palestinian raters judge tokens with markers of dialect, found in less formal speech, to be **less** charismatic than other tokens (r=-.18), but we have no similar annotation for SAE.

($r = .603$). Although the non-native speakers are more conservative in their charisma judgments, they still follow the same trajectory as native speakers in scoring individual tokens. When American (mean: 2.95) and Palestinian (mean: 3.24) raters judge Arabic tokens, the difference in means is also significant, and the normalized ratings of these two groups also show a significant and strong positive correlation (r=.72), with the non-native raters assigning lower charisma scores that are nonetheless correlated with native speaker ratings.

These findings support our examination of individual features and their correlations with the charisma statement, across cultures. In all cases, when different language groups rate the same language their judgments of individual tokens are correlated with one another, even when, as for American and Palestinian ratings of Arabic, and American and Swedish judgments of SAE, the absolute values raters assign to each token are significantly different and more conservative.

6.1. Differences in Ratings Across Cultures

To examine **differences** in ratings between language groups, we identified tokens that elicited significantly different charisma ratings based on the subject's native tongue. Seven Arabic tokens were rated significantly[9] more charismatic by Palestinian subjects, while 1 token was rated significantly less charismatic. American subjects rated 6 SAE tokens as significantly more charismatic than Palestinian subjects, while rating only 1 SAE token as significantly less charismatic. When comparing Swedish and Arabic ratings of SAE tokens, 4 tokens were rated significantly more charismatic by Palestinian subjects and 2 were rated more charismatic by Swedish subjects. While there are too few data points to draw firm conclusions from these differences, we will point out some intriguing trends in the acoustic, prosodic, and lexical features of these 'controversial' tokens. For each pair of subject groups A and B that assessed a set of stimuli, we identify four groups of tokens: those rated 1) significantly less, 2) less, but not significantly less, 3) more, but not significantly more, and 4) significantly more, charismatic by group A than by group B. We examine the mean values of acoustic-prosodic and lexical features for each of these groups, discussing only those features which show monotonic change from token groups 1 to 4.

Arabic tokens rated significantly more charismatic by American subjects than Palestinians tend to have a faster speaking rate and smaller standard deviation of rate over intonational phrases than Arabic tokens rated more charismatic by Palestinians. These tokens also have greater mean intensity and more dynamic intensity[10] than those rated as significantly more charismatic by Palestinians. Palestinians tend to rate Arabic tokens with lower pitch peaks[11] and greater pitch dynamics[12] as significantly more charismatic than Americans do. These differences suggest that Americans find Arabic speakers who employ a faster and more consistent speaking rate, who speak more loudly overall, but who vary this intensity considerably, to be charismatic, while Palestinians show less sensitivity to these qualities. Tokens that Palestinian raters find to be more charismatic than Americans have fewer disfluencies than tokens considered more charismatic by Americans. A more detailed examination of the types of disfluency occurring in each token may help to clarify this finding.

We find fewer instances of monotonic change from groups 1 to 4 when we examine SAE tokens rated significantly differently by Palestinians and Americans. Tokens rated significantly more charismatic by Americans tend to have a higher speaking rate[13] but to be spoken in lower pitch range for the speaker.[14] Note that Americans also find a faster speaking rate more charismatic than Palestinians when they rate Arabic speech.

We now turn our attention to the behavior of Swedish subjects. SAE tokens rated more charismatic by Swedish subjects than by Americans and Palestinians contain speech produced in a more compressed pitch range[15] but with a greater mean (non-normalized) HiF0 value. This high HiF0 value may indicate an expanded pitch range for the speaker, or simply a higher-pitched voice. We also find that Swedish subjects tend to rate tokens with a greater minimum pitch and lower standard deviation of pitch within intonational phrases as more charismatic than Americans. Taken together, these findings suggest that Swedish subjects may find higher pitched speech in a relatively compressed range to be more charismatic than do Americans.

7. Conclusions and Future Research

In this paper we have compared native and non-native raters judgments of speaker charisma from SAE and Arabic stimuli and the acoustic-prosodic and lexical correlates of these judgments. Our findings suggest that, while some acoustic-prosodic correlates are common across cultures, other acoustic-prosodic and lexical correlates are specific to the language rated — yet, curiously, both native and non-native judgments exhibit these correlations. In some cases correlations of rater judgments for speech in their own native language are quite different from correlations of raters from the same language group who judge a foreign language, and resemble charisma correlates of raters from the language being judged. These results are particularly striking for American raters judging Arabic, since these had no knowledge of that language. While, in general, subjects perceive speech in a foreign tongue as less charismatic than tokens in their native language, native and non-native raters tend to agree on the relative charisma of a token, even if they disagree about its absolute rating. Our future work will investigate additional language groups and additional potential correlates of charisma judgments.

8. References

[1] M. Weber, *The Theory of Social and Economic Organization*. OUP, 1947.

[2] P. Boss, "Essential attributes of charisma," *S. Speech Comm. J.*, 41(3), 1976.

[3] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," *EUROSPEECH05*, 2005.

[4] R. Dowis, *The Lost Art of the Great Speech*. New York, 2000.

[5] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psych. Bull.*, vol. 70, pp. 213–220, 1968.

[6] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int'l*, 5(9-10), 2001.

[7] K. Silverman et al., 'ToBI: a standard for labeling English prosody," *ISCLP92*, 1992.

[8] F. Biadsy et al, "Comparing American and Palestinian Perceptions of Charisma Using Acoustic-Prosodic and Lexical Analysis." *INTERSPEECH07*, 2007.

[9] Significance was determined by the Mann-Whitney U test.

[10] Standard deviation of mean intensity over intonation phrases.

[11] Lower maximum pitch and lower maximum HiF0 value.

[12] Standard deviation of pitch over intonational phrases.

[13] Ratio of voiced to unvoiced frames.

[14] Speaker-normalized maximum pitch

[15] Speaker-normalized maximum pitch