INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Prosodic analysis of attention-drawing speech

Carlos Ishi[1], *Jun Arai*[1], *Norihiro Hagita*[2]

[1]ATR/HIL, Japan
[2]ATR/IRC, Japan

carlos@atr.jp, araijun@hotmail.com, hagita@atr.jp

Abstract

The term "attention drawing" refers to the action of sellers who call out to get the attention of people passing by in front of their stores or shops to invite them inside to buy or sample products. Since the speaking styles exhibited in such attention-drawing speech are clearly different from conversational speech, in this study, we focused on prosodic analyses of attention-drawing speech and collected the speech data of multiple people with previous attention-drawing experience by simulating several situations. We then investigated the effects of several factors, including background noise, interaction phases, and shop categories on the prosodic features of attention-drawing utterances. Analysis results indicate that compared to dialogue interaction utterances, attention-drawing utterances usually have higher power, higher mean F0s, smaller F0 ranges, and do not drop at the end of sentences, regardless of the presence or absence of background noise. Analysis of sentence-final syllable intonation indicates the presence of lengthened flat or rising tones in attention-drawing utterances.

Index terms: attention-drawing speech, prosody, intonation, paralinguistic information, expressive speech

1. Introduction

In our definition, the term "attention drawing" refers to the speaking actions of sellers or shopkeepers who attract the attention of people passing by in front of their business spaces (such as shops, restaurants, markets, and theaters) to invite them into their shops.

The background of our research is the development of robots that can socially interact with humans [1-2]. Robots are expected to perform various tasks in stores, including greeting visitors, providing information, and advertising products [3]. Some robots like Pepper (http://www.softbank.jp/en/robot/) are already being deployed for commercial-based services. We aim to develop social robots that can talk and behave like humans. However, it remains challenging to properly control different speaking styles and match them with different situations.

Although the speaking styles of attention-drawing speech are clearly different from conversational speech, few studies have investigated its changes in speaking style. Speakers alter how they produce speech based on the communicative situation. Changes are made to enhance the information transmission's efficiency. For instance, in noisy environments, people speak loudly and produce more energy at higher frequencies (the Lombard effect [4]). Other examples of widely studied differences in speaking styles are read vs. spontaneous speech, and infant-directed vs. adult-directed speech [5-6].

The prosodic features of attention drawing are thought to be related to the distance between the speaker and the listener. The ability to alter speech intensity with changes in listener distance is an important aspect of natural communication. On theory, speech intensity obeys an inverse square law with distance [7]. That is, when the distance between the speakers is doubled, there is a corresponding 6dB reduction in the speech volume due to sound propagation losses. It has been reported that speakers make prosodic, pragmatic, and semantic changes in addition to increasing speech volume to accommodate changes in listener distance [8]. These compensatory changes closely resemble the Lombard effects, which explain that speech intensity is adjusted to compensate for increases in the background noise. Other studies have found that loud speech is also associated with a reduction in speech rate [9].

However, factors other than distance-related changes are thought to be related to attention-drawing speech, since attitudinal inviting or requesting behaviors are involved.

The speaking style of attention drawing may also be related to culture. For example, Sadanobu et al. used the term "street seller's voice" to refer to a category of voice quality that can only be uttered by some (but not all) young Japanese girls, especially if they are selling a product that is deemed cute and/or fashionable [10]. They claim that this voice can be heard at western-style cake shops, but never at "wagashi" shops, i.e., traditional Japanese-style cake shops, since "wagashi" is not cute or fashionable. Their results on three female speakers suggest that a street seller's voice has a "twang" (sharp) quality, which is manifested in the acoustic signal by (among other things) sustained high energy in the upper frequency regions.

In this study, our analysis focuses on the prosodic aspects of attention-drawing speech by considering several of the above factors. We collected the speech data of multiple people with previous attention-drawing experience by simulating several situations and investigated the effects of several factors, including background noise, interaction phases, and shop categories on the prosodic features of attention-drawing utterances.

2. Data collection and annotation

2.1. Data collection

We collected data in our laboratory by recruiting people with experience in attention-drawing and asked them to simulate some attention-drawing situations. We recruited 18 speakers as participants (11 males and 7 females whose ages range from 18 to 43) with previous attention-drawing experience. We also recruited another 4 speakers (2 males and 2 females from 20 to 65) who acted as passersby/customers.

http://dx.doi.org/10.21437/Interspeech.2017-623

The shopkeeper-role participants wore a headset microphone (DPA-4060) and tried to draw the attention of passersby based on their previous experiences and behaved as if they were in front of their stores.

The passerby-role participants were instructed to act based on two conditions. In the first condition, they walked past the shopkeeper and ignored the attention-drawing attempts. In the second, they stopped and reacted to the attention-drawing utterances and asked questions of (or engaged in short interactions with) the shopkeeper based on the shop situation. The conditions for the numbers of passersby were set to a single person, a couple, or a group of four people.

To verify the effects of different noise conditions, we used a pre-recorded shopping mall noise signal to simulate a noisy environment. In the "clean" condition, the background noise level was around 48dBA due to the room's air conditioner. In the "noisy" condition, the noise signal was played through a loudspeaker (Yamaha MS-101III) and the noise level was adjusted to around 65dBA.

The shopkeeper-role participants adapted their attention-drawing speech based on the noise and passerby situations, interacted with passersby who responded to the attention drawing, and finished the interaction by leading them into the shop if they accepted an invitation. To capture the shopkeeper motions, an RGB-D sensor (Kinect-V2) was set about four meters in front of the shopkeeper position.

After each trial, the passerby-role participants answered a questionnaire about their impressions of the attention-drawing attempts that asked the following questions: "Were you drawn to enter the store?", "Was your impression of the attention-drawing favorable?", and "Was the voice easy to listen to?". We removed the data of three of the male shopkeeper participants who received negative grades, for the analysis.

The following shop categories resulted from our data collection (the speaker IDs in parentheses include the speaker gender (F or M), age, and initials): restaurant street (F22SS), "ozouni" (traditional Japanese New Year's food) at a shrine (M22AK), "wagashi" (Japanese cake) shop (F21HY), "izakaya" (Japanese-style bar/restaurant) (M25ST, M43HH), Christmas cake shop (F24RS), questionnaire inquiry (M21YI), clothing store (M21MH, M22WH, F40MR), food store (M22SY), drugstore (F22AM), lottery event (F18CY), beer selling at October Fest (M23YE), food sale (F30CF), and gift shop (M22MK).

The shopkeeper's utterances (captured by the headset microphone) were segmented and transcribed. The following text patterns were frequently found during attention-drawing: "irasshamase" ("Welcome," "May I help you?"), "konnichiwa" ("Hello"), "ikagadeshouka" ("Would you like"), "gozaimasu" ("We have"), "tabete kudasai" ("Try/taste"), "okaidoku tonatte orimasu" ("It's on sale"), "otameshi kudasai" ("Try"), "hanbai shite orimasu" ("We are selling"), "seeru o okonatteorimasu" ("We have a sale on"), "hanbaichuu desu" ("We are selling"), and "douzo gorankudasai" ("Have a look on").

2.2. Annotation data

From the collected data, we observed the following interaction sequence shared by all the shopkeepers. The sequence starts with an attention-drawing phase, where the shopkeeper's utterances are directed to unspecified people (general public): Phase 1. Then, as passersby approach, the attention-drawing utterances become more directed to specific individuals: Phase 2. When passersby stop and ask the shopkeeper questions, the interaction shifts to a dialogue phase, where utterances are clearly directed to specific individuals: Phase 3. Finally, the interaction finishes with a parting greeting or a guiding utterance into the shop: Phase 4.

Every interaction is segmented by the above interaction phases and annotated according to the following set of labels: {a (attention-drawing phase), d (dialogue phase), f (finalization phase)}, and {p (public-directed), ip (ambiguity between public and individual-directed), and i (individual-directed)}.

The data were annotated by two research assistants and later combined. The agreement rates were 91% for the {a, d, f} label set and 84% for the {p, ip, i} label set. Most of the mismatches occurred in the transitions between public-directed ("p") and individual-directed ("i") utterances. For these utterances, the "ip" label was attributed.

3. Analysis results

We analyzed the effects of several factors, such as interaction phases, presence/absence of noise, numbers of passersby, gender of shopkeepers, and the shop categories on the speaking style of the shopkeepers.

3.1. Speaking styles in different interaction phases

Figure 1 shows examples of the F0 contours of the shopkeepers during one interaction with a passerby. The horizontal dotted lines are plotted for each octave (corresponding to 55, 110, 220, and 440 Hz). For the F0 analysis, all F0 values were converted to a musical scale (log scale) before processing.

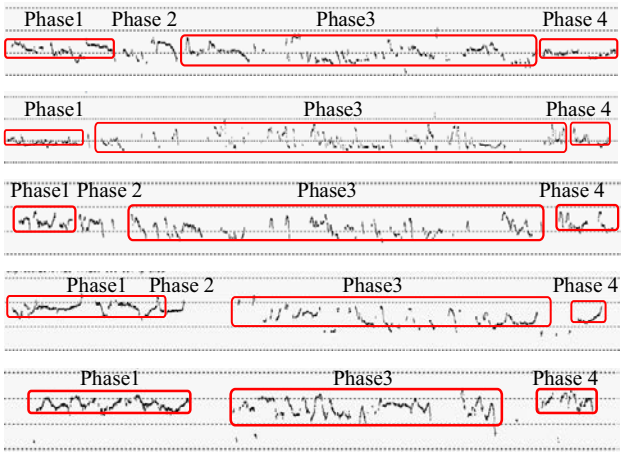

Figure 1: *Examples of F0 contours of shopkeepers during one interaction with a passerby. Red boxes separate different phases during an interaction.*

Common features among the shopkeepers are clearly observed from these examples. In the public-directed attention-drawing phase (Phase 1), the F0 range is narrower, and the F0 contours do not drop at the end of the sentences as in the individual-directed dialogue phase (Phase 3). In the finalization phase (Phase 4), the F0 features draw closer to the attention-drawing phase, depending on the situation. The individual-directed attention drawing phase (Phase2) show intermediate features between Phase 1 and Phase 3. In some of the speakers, Phase 2 is very short or inexistent.

Figure 2 shows the differences for each speaker in the following values between the public-directed attention drawing (Phase 1) and individual-directed dialogue (Phase 3): power average (*Power_avg*), F0 average (*F0_avg*), and F0 standard deviation (*F0_std*). Data of two of the speakers (M25ST and M21YI) are not shown because their utterances were always directed to individuals, i.e. Phase 1 was inexistent. The upper and lower panels respectively show the results for the clean and noisy conditions. In almost all of the speakers, the *F0_avg* was higher (positive *F0 avg diff* values), and the standard deviations were smaller (negative *F0 std diff* values) during the attention-drawing phase. The mean F0s were 1 to 3 semitones higher (p<0.05 by t-test), and the standard deviations were 1 to 2 semitones lower (p<0.01 by t-test) in the attention-drawing phase than the dialogue phase.

Regarding power, the average power was 3 ~ 9 dB larger in the attention-drawing phase (p<0.01 by t-test). The variations in power depended on the speaker and the shop category, but they are less dependent on the noise condition. For example, in the questionnaire inquiry, attention drawing was usually directed politely to a specified person, so the differences in the prosodic features with the dialogue mode became smaller.

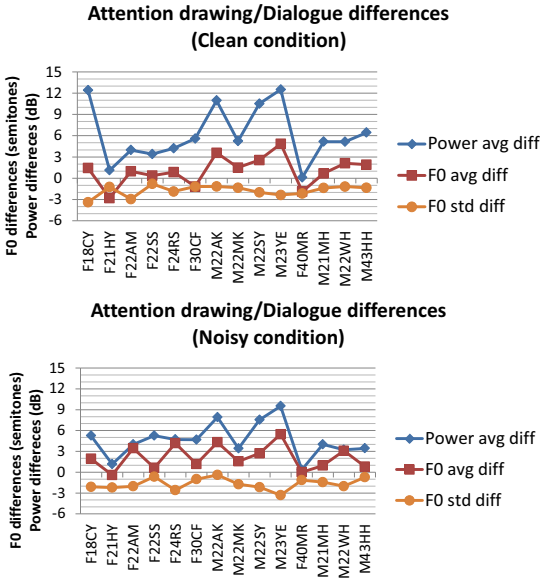

Figure 2: *Differences in power and F0 features between attention-drawing and dialogue phases for each speaker.*

Comparing the results of the clean and noisy conditions, we observed similar trends for the differences in speaking style between the attention-drawing and dialogue phases (all differences were significant with p<0.01 by t-tests). This means that although there are differences in the vocal effort to compensate for the background noise, consistent differences exist in the speaking styles between different interaction phases. In other words, the differences due to the interaction phases are superimposed on the differences due to the background noise discussed in Section 3.2.

The distributions for the individual-directed attention drawing (Phase2) and finalization (Phase 4) utterances were intermediate between public-directed attention drawing (Phase 1) and individual-directed dialogue (Phase 3) utterances.

3.2. Effects of background noise

Regarding the effects of background noise, we found in most of the speakers that power increases, F0 becomes higher, and the speaking rate becomes slower. These changes in the prosodic features are related to the Lombard effect. Fig. 3 shows the differences in power and F0 features between clean and noisy conditions. In both the attention-drawing and dialogue phases, the power differences are around 9dB (p<0.01 by t-test). Higher F0 values (positive *F0 avg diff* values) were observed in almost all the speakers. However, these differences are smaller in the dialogue phase compared to the attention-drawing phase. The average F0s are 1 to 3 semitones higher in the attention-drawing phase (p<0.01 by t-test) and around 0 for some of the speakers in the dialogue phase. An interesting result is that the no significant differences were found in the standard deviation of the F0 values (*F0 std diff* values around 0; (p=0.33 for attention-drawing and p=0.48 for dialogue phase), which means that although vocal efforts were increased to compensate for the background noise, the F0 ranges (in log scale) did not change.

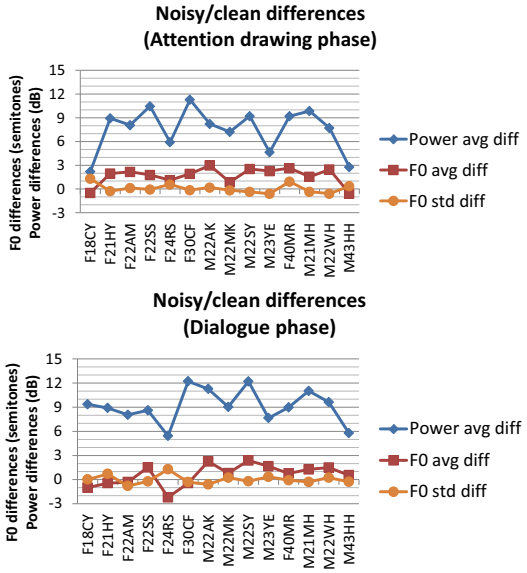

Figure 3: *Differences in power and F0 features between clean and noisy conditions for each speaker.*

3.3. Sentence-final intonation

From a preliminary qualitative analysis, we found that sentence-final intonation tends to lengthen with a rising tone in attention-drawing utterances. We then segmented the sentence's last syllable and analyzed the sentence-final intonation. We focused our analyses on sentence types that frequently occurred in the attention-drawing utterances. The following sentence types were frequently found (the numbers in parentheses are the total of the unspecified people-directed utterances for male and female speakers): "irasshamase" (146+131), "konnichiwa" (32+56), "douzo" (41+26), "ikagadesuka" (36+19), "ikagadeshouka" (31+47), "kudasai" (1+12), and "…masu" (155+118). The total number of unspecified people-directed utterances was 166 for males and 79 for females.

Figure 4 shows the distributions of the sentence-final tones of attention-drawing utterances in the public-directed (Phase 1) and individual-directed (Phase 2) utterances for male and

female speakers. The vertical axis in Figure 4 represents the duration of the sentence-final syllables, and the horizontal axis represents *F0move*, a parameter that quantifies the F0 movement within the last syllable. *F0move* is calculated as the difference between the target F0 value of the first-order regression line at the syllable's end and the average F0 value of its first half [11-12]. *F0move* values larger than 1 semitone indicate rising tones, -2 ~ 1 semitones indicate flat tones, and smaller than -2 semitones indicate falling tones.

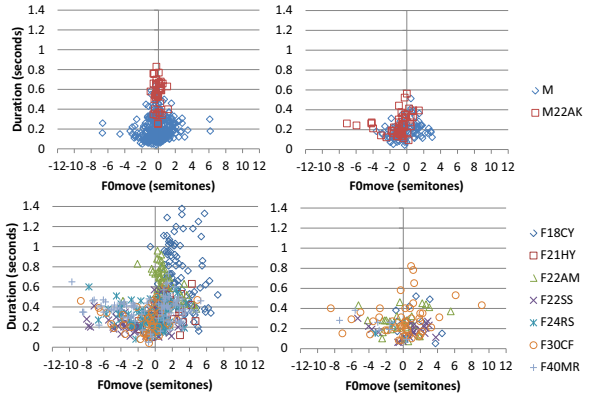

Figure 4: *Sentence-final tones (F0move vs. duration) for public-directed (left panels) and individual-directed (right panels) attention-drawing utterances in male (upper panels) and female (lower panels) speakers.*

The results of the public-directed utterances in the left panels of Figure 4 suggest that a wider range of intonation styles is used by female than male speakers. For male speakers, short flat tones and short rising tones are predominant (0.21±0.11 seconds; 0±1.5 semitones). In speaker M22AK, long flat tones (0.55±0.14 seconds; 0±0.4 semitones) are predominant. In female speakers, sentence-finals that were lengthened to more than 0.5 seconds were found for some of the speakers (0.69±0.3 seconds for F18CY and 0.64±0.15 seconds for F22AM). A clear feature was found that they were either flat or rising tones (2.5±1.9 semitones for F18CY and 1.1±1.5 semitones for F22AM). Falling tones were mostly found when the duration was shorter than 0.5 seconds.

For individual-directed utterances (right panels of Figure 4), for both male and female speakers, the sentence-finals were generally not lengthened (0.18±0.08 seconds for male speakers and 0.24±0.14 seconds for female speakers). Short flat and short rising tones are predominant.

4. Discussion

In this section, we discuss about the results of the speakers who showed different trends in the analyses of Section 3, and present some results concerning other factors.

In some of the speakers (F18CY, M23YE, and M43HH), the average power differences were smaller between the noisy and clean conditions for the attention-drawing utterances (Fig. 3). This was because these speakers were already speaking loudly even in the clean conditions. The upper panel of Fig. 2 shows that the differences in power are higher for these speakers. On the other hand, speaker F40MH showed fewer differences in her average features between attention-drawing and dialogue modes (Fig. 2). This was because she also talked in a relatively loud voice during the dialogue phase and often laughed while speaking, increasing both her power and F0.

Regarding the shop category factor, there might be some relationship with gender or stereotypes. For example, the speaker F21HY, who works in a "wagashi" shop, was the only speaker who showed lower average F0s during the attention-drawing phase (negative *F0avg diff* values in Fig. 2). This might be related to the fact that such traditional shops try to present a dignified atmosphere, and so she avoids pitch increases like in western cake shops and supermarkets [10].

From the analyses of sentence-final intonation, we found that female speakers have a wider range of intonation, including long rising tones, than male speakers (Fig. 4). It is known that sentence-final rising intonation can occur in several situations, including yes-no questions, information requests, agreement requests, opinion requests, and proposals/invitations [13]. Furthermore, a long rise intonation may evoke a gentle style [14]. The presence of long rising tones (including declarative sentences) in public-directed attention drawing by female speakers (right bottom panel of Fig. 4) might be related to a gentle manner to make a request or an invitation.

Regarding different conditions of the passersby (number, gender and age), some of the shopkeepers changed their utterance contents according to the gender and age of the passersby. However, no clear differences were found in the speaking styles of the attention-drawing utterances between these conditions.

Finally, regarding the distance between the shopkeeper and the passersby, as a general trend, the human positions estimated from the Kinect sensor data indicated that when the distance closed within around two meters, the shopkeepers switched their attention-drawing style from public-directed to individual-directed. As stated in the introduction, increases in power and F0 might be partly due to compensations of the distance between speakers and listeners. However, the other features of smaller F0 ranges and lengthened flat or rising sentence-final tones can be considered to be specific of public-directed attention-drawing styles.

5. Conclusions

We analyzed the prosodic features of the attention-drawing speech of several speakers with experience in attention-drawing and investigated the effects of several factors, including background noise, interaction phases, and the shop categories. Analysis results indicated that attention-drawing utterances usually have higher power, higher mean F0s, smaller F0 ranges, and do not drop at the end of sentences than dialogue interaction utterances, regardless of the presence or absence of background noise. Analysis of sentence-final syllable intonation indicated that long flat or long rising tones appear with higher frequency in public-directed attention-drawing utterances than in individual-directed ones.

Future works include the analysis of gestures and the incorporation of the rules inferred from the present analysis to the prosody control of a robot's voice.

6. Acknowledgements

This work was partly supported by JST/ERATO (Grant Number JPMJER1401) and MIC/SCOPE. We thank Taeko Murase, Kyoko Nakanishi and Miki Okuno for contributions in the data analysis.

7. References

[1] Glas, D.F., Minato, T., Ishi, C.T., Kawahara, T., & Ishiguro, H. "ERICA: The ERATO Intelligent Conversational Android," Proc. of *25th IEEE International Symposium on Robot and Human Interactive Communication* (*RO-MAN 2016*), pp. 22-29, Aug. 2016.

[2] Oishi, Y., Kanda, T., Kanbara, M., Satake, S., & Hagita, N., "Toward End-User Programming for Robots in Stores," Proc. of the *2017 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI2017*), pp. 233-234, 2017.

[3] Gross, H.-M., et al., "Shopbot: Progress in Developing an Interactive Mobile Shopping Assistant for Everyday Use," *SMC2008*, pp. 3471-3478, 2008.

[4] Junqua, J.C. The Lombard reflex and its role on human listeners and automatic speech recognizers. *J. Acoust. Soc. Am. 93*, 510-524 (1993)

[5] Garnica, O.K. (1977). "Some prosodic and paralinguistic features of speech to young children," In C.E.Snow & C.A.Ferguson (Eds.), *Talking to Children*, pp. 63-88. Cambridge: Cambridge University Press.

[6] Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A crosslanguage study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language 16*, 477-501, 1989.

[7] Zahorik, P. & Kelly, J.W. (2007). Accurate vocal compensation for sound intensity loss with increasing distance in natural environments. *Journal of the Acoustical Society of America 122*(5), 143-150, 2007.

[8] Michael, D.D., Siegel, G.M., & Pick, H.L. (1995). Effects of distance on vocal intensity. *Journal of Speech and Hearing Research, 38*, 1176-1183, 1995.

[9] Schulman, R. (1989). Articulatory dynamics of loud and normal speech. *Journal of the Acoustical Society of America 85*(1), 295-312, 1989.

[10] Sadanobu, T., Zhu, C., Erickson, D., Obert, K. "Japanese "street seller's voice"," Proc. *172nd Meeting of the Acoustical Society of America*, Paper 5aSC46, 2016.

[11] Ishi, C.T., Ishiguro, H., Hagita, N. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication 50*(6), 531-543, June 2008.

[12] Ishi, C.T. (2006), The functions of phrase final tones in Japanese: Focus on turn-taking. Journal of Phonetic Society of Japan, Vol. 10 No.3, 18-28, Dec. 2006.

[13] Hatano, H., Kiso, M., & Ishi, C. "Analysis of factors involved in the choice of rising or non-rising intonation in question utterances appearing in conversational speech," Proc. *14th Annual Conference of the International Speech Communication Association* (*Interspeech 2013*), August, 2013, pp. 2564-2568.

[14] Ishi, C. T., "Perceptually-related F0 parameters for Automatic Classification of Phrase Final Tones", *IEICE Trans. Inf. & Syst. E88-D* (3), 481-488, 2005.