Journal of the Korea Convergence Society
Vol. 11. No. 9, pp. 91-97, 2020

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2020.11.9.091

Research on Emotional Factors and Voice Trend by Country to be considered in Designing AI's Voice
- An analysis of interview with experts in Finland and Norway

Kiechan Namkung
Professor, Industry Academic Cooperation Foundation, Kookmin University

AI의 음성 디자인에서 고려해야 할 감성적 요소 및 국가별 음성 트랜드에 관한 연구
- 핀란드와 노르웨이의 전문가 인뎁스 인터뷰를 중심으로

남궁기찬
국민대학교 산학협력단 교수

Abstract Use of voice-based interfaces that can interact with users is increasing as AI technology develops. To date, however, most of the research on voice-based interfaces has been technical in nature, focused on areas such as improving the accuracy of speech recognition. Thus, the voice of most voice-based interfaces is uniform and does not provide users with differentiated sensibilities. The purpose of this study is to add a emotional factor suitable for the AI interface. To this end, we have derived emotional factors that should be considered in designing voice interface. In addition, we looked at voice trends that differed from country to country. For this study, we conducted interviews with voice industry experts from Finland and Norway, countries that use their own independent languages.

Key Words : AI, Voice user interface, Emotional factor, Voice trend, Voice identity

요 약 사용자와의 인터랙션이 가능한 음성 기반의 인터페이스는 AI 기술의 발달에 따라 사용이 확대되고 있다. 하지만, 현재까지의 음성 기반 인터페이스에 대한 연구는 음성 인식의 정확성 향상 등 기술적인 연구들이 대부분이었다. 이렇다 보니, 대부분의 음성 기반 인터페이스의 목소리는 차별화된 감성을 제공하지 않으며 획일화되어 있다. 본 연구에서는 AI 인터페이스의 음성에 적합한 감성 요소를 더하는 것을 목적으로 한다. 이를 위해 음성 인터페이스 디자인에서 고려되어야 할 감성적 요소를 도출하였다. 또한, 국가별로 차이를 보이는 보이스 트렌드를 조사하였다. 본 연구를 위해 자국의 언어를 독립적으로 사용하는 핀란드와 노르웨이, 두 국가의 음성 산업 전문가들과 인터뷰를 진행하였다.

주제어 : 인공지능, 음성 인터페이스, 감성 요소, 음성 트랜드, 음성 아이덴티티

*Corresponding Author : Kiechan Namkung(Kookmin University)
Received June 27, 2020 Revised July 30, 2020
Accepted September 20, 2020 Published September 28, 2020

1. Introduction

Voice feedback of devices with text-to-speech (TTS) technology has recently evolved into voice user interfaces (VUIs) that enable human-computer interaction following the development of AI technology in the fourth industrial revolution. In 2011, Apple introduced its virtual assistant, Siri, a voice interface that allows users to communicate with AI on mobile phones; since then, global companies such as Google, Microsoft, and Samsung have competitively installed their own proprietary interactive voice assistants on a variety of AI devices. This voice-based AI interface will certainly be a key technology in the near future, its applications ranging from self-driving cars to smart homes and beyond; hence, VUI-related research is becoming increasingly important not only in industry but also in academia.

To date, most of the research on voice-based AI interfaces has been technical studies focused on improving the accuracy of speech recognition. These studies have put forward new solutions for improving speech recognition accuracy through experiment and verification and also studied the improvement of speech recognition accuracy in various spaces [1-4]. While fewer in number relative to these technology-based studies, there are also studies that address topics related to the human factors of a voice-based interface. Recently, several studies have been conducted regarding the gender of voice in voice-based interfaces[5-7].

A study on the gender of a device-mounted voice assistant provided a cultural rationale for designing a voice assistant with a generalized female voice [5]. Another study addressed differences in user perception in relation to gender in the voice interface[6], and yet another examined the relationship between the user's gender and their gender preference for the voice assistant[7].

Apple offers 41 languages for its voice assistant, Siri (based on iOS 12.3.1, July 2019). Of these, four of the languages are spoken in eight countries—Dutch, Arabic, British English, and French—and are provided by default with a male voice, while the languages of the other 33 countries are provided with a female voice by default. However, even if the default voice is female, users of most languages have the option to change them into male voices.

Why, then, does Apple offer both male and female voices on a single device? Why did Apple define the male voice as the default for languages in only eight countries? We inferred that both male and female voices are provided on one device in consideration of various user preferences. There are, however, several reasons for why a male voice was defined as the default in only eight languages. If we construct a synthesized voice from a real voice, the degree of naturalness may vary depending on the linguistic characteristics of that language and the sex and tone of the voice. Aside from these technical problems, there may have been unexpected problems such as issues related to the schedule of the voice actor. Perhaps it was a decision that considered the preferences of users in each region. For example, users in Singapore, Britain, and India who speak British English may prefer a male voice.

In any case, the voice of most voice-based interfaces may be male or female. However, there is no study that suggests that limiting users' voice choices to gender only is the answer.

As in the case of Siri, in order to produce a voice interface that is usable in a variety of languages, we clearly need to consider more complex issues than we initially thought. From a user-centered design perspective, companies that develop voice interfaces must be cognizant of the preferred speech characteristics of users in various countries and consider such preferences when determining the default value of the device.

The characteristics of a user's preferred voice may be its tone or its gender. These factors will be based on the emotions evoked in users as they listen to the voice of the interface, which will result in differences related to the cultural characteristics of each country.

Therefore, this study seeks to derive emotional factors of judgment to select the voice of the voice-based interface from a user-centered design perspective and to determine whether there is a voice preference due to cultural differences in each country. To this end, we interviewed experts in the voice industry in Finland and Norway, as representatives of countries that use their own languages.

2. Literature review

2.1 Gendered characteristic of voice-based interface

Most of the voice assistants that are now released have gendered features. Apple's Siri, which means "the beautiful woman who leads the victory" in Norwegian, was first released as a female voice, while Google's Assist offers a female voice as a default, and Microsoft's Cortana is named after a female character in the popular video game, Halo. When questioned about the genders of these voice assistants, companies such as Amazon and Microsoft have said they developed products based on consumer responses, which indicated that users prefer to be clearly aware of the gender of the assistant's voice and feel more comfortable with a female voice than with a male one [5].

Using voice feedback from a device that features the voice of a woman is not a new practice. Since the time when interactive voice response (IVR) was first developed and used, most machines have adopted a female voice[8]. This tendency prevailed because female voices were better heard and hence favored by consumers. However, some scholars have pointed out that this reason is prejudiced[9]. Currently, "Equal AI Initiative,"a group that raises awareness of sexual inequality in artificial intelligence technology, is active in the U.S.[10].

Gender may be an important factor in the choice of speaker in a voice-based interface. However, the reason for the gender selection should not simply be based on the vague belief that it sounds good and consumers will like it.

2.2 Emotional factors that determine the characteristics of a voice

When people deliver messages, the weight for delivery is 38 percent for voice, 55 percent for nonverbal elements, and only 7 percent for speech [11]. We can interpret this spread to mean that a "good" voice makes one-third of the message delivery already successful. Hence, the selection of voice used for voice-based interfaces that send messages to users is a critical factor in the delivery of information.

The voice of announcers who communicate factual information to people is considered a very important social asset. A study that analyzed the voices of such announcers identified certain vocal elements (i.e., pitch, energy carried on the voice, frequency variation, amplitude variation, and speech speed) as factors that can affect the reliability and stability of the news that listeners feel[12].

In a study exploring the correlation between characteristics of speech and their effect on sales, the characteristic factors of speech were divided into speaking rate, fundamental frequency, and loudness variability[13].

Another study divided the elements of voice characteristics into age, cheerfulness, sternness, gender, and speaking rate and studied their effects on speech recognition[14]. In studies that related voice interfaces to social identity theory (SIT)

using the method of reliability evaluation[15] and the method of measurement of individual attraction [16], voice characteristics used in voice interfaces were evaluated as factors such as credibility, social attraction, and task attraction[17].

Although the emotional factors of speech used in these preceding studies differ in name, it can be seen that many of the factors actually have a common meaning.

3. Method

3.1 Time period

The interviews were conducted in May 2018 in Helsinki, Finland, and Oslo, Norway.

3.2 Participants

Ten voice industry experts from Finland and Norway participated in the interview.

Participants in Finland included two dubbing producers, one voice actor, one project manager of a media company, and one creative director of an audio studio, and in Norway, one dubbing director, one sound director, one sound art director, one voice casting manager, and one voice actor. All participants are experts in the voice industry with 10 to 40 years of experience.

3.3 Procedures

We interviewed each expert for half an hour. Before the interview, they tried to use Apple's Siri, a voice assistant offered in their own language, and recognized the tone and status of the voice assistant. In the interview, we listened to their opinions on emotional factors in designing the voice for the voice-based AI interface. Then, we asked them country-specific questions about characteristics related to voice trends in the local voice industry, the voice preferences of users in each country and why such preferences existed.

Table 1. Experts who took part in the interview

Country	Participant (Occupation)	Career
Finland	A (Dubbing producer)	Has worked in the voice industry since 1991. Works in Finland's largest dubbing studio. More than 250 movie dubbing works.
	B (Dubbing producer)	Has worked as a dubbing producer since 1994. Advertising and film dubbing.
	C (Voice actor)	Has worked as a voice actor since 1975. Famous voice in Finland.
	D (Project manager)	Has worked in the voice industry since 1999. In charge of producing voices in the works of big retailers such as Disney and Warner.
	E (Creative director)	Has worked in the voice industry since 1992. Sound Creative Director for Finnish Radio.
Norway	A (Dubbing director)	Works at Oslo's largest dubbing studio. Creates dubbing works such as animations, documentaries, advertisements, computer games, etc.
	B (Sound director)	Responsible for sound work on voice dubbing.
	C (Sound art director)	Art director of advertising company, responsible for voice work used in radio and TV advertising
	D (Voice casting manager)	Voice casting used in animation, movies, etc.
	E (Voice actor)	Dubbed voice of a famous Norwegian film.

4. Findings

4.1 Emotional factors to be considered in voice design

In the interviews, the experts gave the following emotional factors as characteristics to consider when designing the voice for a voice-based AI interface (as shown in Table 2).

Although the experts from the two countries presented various opinions, we could see an overlap of expressions of similar meaning. We classified the emotional factors presented by the 10 experts into similar attributes using the Affinity Diagram, a user-experience design methodology. As a result, we divided the factors into three representative groups: Dynamic, Inviting, and Trustworthy, as shown in Table 3.

Table 2. Emotional factors to be considered in designing AI's voice

Country	Participant	Emotional Factors
Finland	A	voice of intimacy, kindness, voice interested in conversation, not-dull voice, positive and bright voice, warm, reliable voice.
	B	smiley voice, natural and stable voice, comfortable voice.
	C	fatherly voice, balanced voice, unexaggerated voice, confident voice, authoritative voice.
	D	voice with a positive smile, voice with a proper rate of ignition,natural voice, placid voice.
	E	Has worked in the voice industry since 1992. Sound Creative Director for Finnish Radio.
Norway	A	comfortable voice, neutral voice, low voice, trustworthy voice.
	B	trustworthy voice, the voice of an adult, friendly voice.
	C	warm voice, comfortable voice, natural voice, friendly voice.
	D	neutral voice, easily tireless voice, calm voice, confident voice, dynamic voice.
	E	voice without female-specific nagging, pleasant, relaxed voice, distinctive voice, moderately lively voice

These three representative factors include not only the opinions of experts, but also the elements used in the preceding study, which dealt with the characteristics of the voice discussed earlier.

Table 3. Three representative emotional factors derived by bottom-up classification method

Emotional factors	Detail elements
Dynamic	interested in conversation, not-dull, bright, proper rate of ignition, low, tireless, dynamic, without female-specific nagging, moderately lively
Inviting	Intimacy, kindness, warm, smiley, natural, comfortable, positive smile, placid, jolly, friendly, warm, calm, pleasant, relaxed, distinctive
Trustworthy	positive, reliable, stable, fatherly, balanced, unexaggerated, confident, authoritative, neutral, clever, trustworthy, adult, confident

4.2 Voice trends by country

In the interviews, we asked the 10 voice industry experts from the two countries about their local voice trends. A summary of the interviews is shown in Table 4.

Table 4. Voice trends by country

Country	Voice Trends
Finland	Traditionally, Finnish people prefer low-pitched, mature male voices. Finnish people prefer a voice that can convey feelings of trust and power. Trends are changing to a natural and youthful feel. In the media, voices that are unique and can convey local color are also popular. There seems to be no big difference between regions within Finland. In Finland, a feeling of trust is the most important factor.
Norway	Local media are legally required to air a mixture of 20% dialects. Dialect has a great influence on the public. Men-on-the-street voice is popular. A unique voice is important, with a preference for a "young" voice. Low voices are traditionally always popular. There is no preference for gender of voice.

Regarding the voice trend in Finland, Finnish people traditionally prefer a low-pitched male voice that can give them confidence and trust. They also said that there was no particular preference for gender, and that current trends tend to prefer young-feeling voices.

In Norway, the law stipulates that the media in each region use 20% dialect. As such, they said that individuality of the voice is very important. Norwegians traditionally tend to prefer low voices, and as in Finland, there is no particular preference for gender.

4.3 Weight of emotional factors

We asked voice industry experts from the two countries who participated in the interview about the importance of emotional factors to be considered in the design of voice used in voice interfaces in each of their languages. Experts responded to the importance of each of the three emotional factors, so that the sum of the three factors of Dynamic, Inviting, and Trustworthy was 100%. Through this survey, we derived the importance of and differences in emotional factors to be considered in the design of voice interfaces that consider voice trends in each country.

According to the survey, in Finland, Trustworthy had 46%, Dynamic, 28%, and Inviting, 26%, In Norway, Inviting received 40%, Trustworthy, 32%, and Dynamic, 28%. These results reflect Finland's voice trend, which traditionally values low-pitched, reliable voices, and that of Norway, which values regional color and individuality of voice.

5. Conclusion and Discussion

The results of expert interview analysis in Finland and Norway are as follows:

First, the three emotional factors of Dynamic, Inviting, and Trustworthy should be considered when designing the voice for an AI-based voice interface. Experts have mentioned dozens of emotional factors to be considered in designing AI voice and these three factors represent most of the emotions that people might feel when using voice-based interfaces.

Second, the importance of these emotional factors may vary from country to country. According to our results, Trustworthy was the most important factor for designing voices in Finland, whereas Inviting was the most important factor for designing voices in Norway, where unique voices are important.

Third, when designing the voice of a voice-based interface, ensuring that it matches the purpose and context of the product seems to be more important than simply choosing a voice of a preferred gender. In Finland, the voice trend has traditionally favored low-to-middle pitched male voices, but experts stated that it was more important to choose a gender that fits the purpose and situation of the product's use than this voice trend. Sufficiently considered choice of voice gender can lead users to understand gender issues that have recently emerged.

Fourth, different countries have their own voice trends, and voice-based interfaces produced in consideration of these voice trends can be more appealing to users.

Voice-based interfaces are increasingly widespread in their use. The current research on the development of voice interfaces has focused on technical importance, such as the correct recognition of user ignition and natural voice feedback to suit the situation. However, in the future, it seems it will be necessary to take a more emotional approach, accounting for factors such as the feeling of a voice, voice trends that reflect the cultural characteristics of each language-speaking country, and voice preferences of users. Products with a voice-based interface that reflects this approach may be a big differentiator from those that do not, and the company that incorporates these findings may have the opportunity to have its own voice identity in a uniform voice-based interface market.

REFERENCES

[1] Homma, T., Obuchi, Y., Shima, K., Ikeshita, R., Kokubo, H. & Matsumoto, T. (2018). In-vehicle voice interface with improved utterance classification accuracy using off-the-shelf cloud speech recognizer. *IEICE Transactions on Information and Systems, E101D(12),* 3123-3137.
DOI : 10.1587/transinf.2018EDK0001.

[2] Scanion, P., Elliss, D. P. W. & Reilly, R. B. (2007). Using broad phonetic group experts for improved speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 15(3),* 803-802.
DOI : 10.1109/TASL.2006.885907.

[3] Ramirez, J., Segura, J. C., Gorriz, J. M. & Garcia, L. (2007). Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. *IEEE Transactions on Audio, Speech & Language Processing, 15(8),* 2177-2189.
DOI : 10.1109/TASL.2007.903937.

[4] Zolnay, A., Kocharov, D., Schluter, R. & Ney, H. (2007). Using multiple acoustic feature sets for speech recognition. *Speech Communication, 49(6),* 514-525.
DOI : 10.1016/j.specom.2007.04.005.

[5] HeeEun, L. (2018). Why do voice-activated technologies sound female? Sound technology and gendered voice of digital voice assistants. *Korean*

Journal of Communication & Information, 90, 126-153.

[6] Nguyen, Q. N., Ta, A. & Prybutok, V. (2019). An integrated model of voice-user interface continuance intention: The gender effect. I*nternational Journal of Human-Computer Interaction, 35(15),* 1362-1377.
DOI : 10.1080/10447318.2018.1525023.

[7] Mabanza, N. (2018, December). Gender influences on preference of pedagogical interface agents. *Proceedings of the International Conference on Intelligent & Innovative Computing Applications*, Plaine Magnien, Mauritius.
DOI : 10.1109/ICONIC.2018.8601292.

[8] Couper, M. P., Singer, E. & Tourangeau, R. (2004). Does voice matter? An interactive voice response (IVR) *experiment. Journal of Official Statistics, 20(3),* 551-570.

[9] Myles, J. F. (2013). Instrumentalizing voice: Applying Bahktin and Bourdieu to analyze interactive voice response service. *Journal of Communication Inquiry, 37(3),* 233-248.
DOI : 10.1177/0196859913491765.

[10] Sydell, L. (2018). *The push for a gender-neutral Siri*. [Online].
www.npr.org/2018/07/09/627266501/the-push-for-a-gender-neutral-siri

[11] Mehrabian, A. & Ferris, S. R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology, 31(3),* 248-252.

[12] GwanHae, C., DongUk, C., BumJoo, L., Yeong, P. & YeonMan, J. (2017). A study on characterizing the voices of active announcers using voice analysis technology. *The Journal of Korean Institute of Communication and Information Sciences, 42(7),* 1422-1431.
DOI : 10.7840/kics.2017.42.7.1422.

[13] Peterson, R. A., Cannito, M. P. & Brown, S. P. (1995). An exploratory investigation of voice characteristics and selling effectiveness. *Journal of Personal Selling & Sales Management, 15(1),* 1-15.

[14] Suzuki, H., Zen, H., Nankaku, Y., Miyajima, C., Tokuda, K. & Kitamura, T. (2003). *Speech recognition using voice-characteristic-dependent acoustic models. Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China.
DOI : 10.1109/ICASSP.2003.1198887.

[15] McCroskey, J. C. & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs, 66,* 90-103.
DOI : 10.1080/03637759909376464.

[16] McCroskey, J. C. & McCain, T. A. (1974). The measurement of interpersonal attraction. *Speech Monographs, 41,* 261-266.
DOI : 10.1080/03637757409375845.

[17] Chad, E., Automn, E., Brett, S., Xialing, L. & Noelle, M. (2019). Evaluations of an artificial intelligence instructor's voice: *Social identity theory in human-robot interactions. Computers in Human Behavior, 90,* 357-362.
DOI : 10.1016/j.chb.2018.08.027.

남궁 기 찬(Kiechan Namkung) [정회원]

· 2019년 2월 : 국민대학교 경험디자인과(디자인학박사)
· 2020년 9월 ~ 현재 : 국민대학교 산학협력단 전임연구교수
· 관심분야 : UX, AUI, VUI
· E-Mail : kc.namkung@gmail.com