Journal of The Korea Society of Computer and Information
Vol. 19, No. 11, November 2014

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2014.19.11.159

감정 표현 방법: 운율과 음질의 역할

이 상 민*, 이 호 준**

How to Express Emotion: Role of Prosody and Voice Quality Parameters

Sang-Min Lee *, Ho-Joon Lee **

요 약

본 논문에서는 감정을 통해 단어의 의미가 변화될 때 운율과 음질로 표현되는 음향 요소가 어떠한 역할을 하는지 분석한다. 이를 위해 6명의 발화자에 의해 5가지 감정 상태로 표현된 60개의 데이터를 이용하여 감정에 따른 운율과 음질의 변화를 살펴본다. 감정에 따른 운율과 음질의 변화를 찾기 위해 8개의 음향 요소를 분석하였으며, 각 감정 상태를 표현하는 주요한 요소를 판별 해석을 통해 통계적으로 분석한다. 그 결과 화남의 감정은 음의 세기 및 2차 포먼트 대역너비와 깊은 연관이 있음을 확인할 수 있었고, 기쁨의 감정은 2차와 3차 포먼트 값 및 음의 세기와 연관이 있으며, 슬픔은 음질 보다는 주로 음의 세기와 높낮이 정보에 영향을 받는 것을 확인할 수 있었으며, 공포는 음의 높낮이와 2차 포먼트 값 및 그 대역너비와 깊은 관계가 있음을 알 수 있었다. 이러한 결과는 감정 음성 인식 시스템뿐만 아니라, 감정 음성 합성 시스템에서도 적극 활용될 수 있을 것으로 예상된다.

▸ Keywords : 감정 음향 요소, 감정 운율 구조, 감정 음질 변수, 한국어 감정 음성 분석, 감정 음성 인식, 감정 음성 합성

Abstract

In this paper, we examine the role of emotional acoustic cues including both prosody and voice quality parameters for the modification of a word sense. For the extraction of prosody parameters and voice quality parameters, we used 60 pieces of speech data spoken by six speakers with five different emotional states. We analyzed eight different emotional acoustic cues, and used a discriminant analysis technique in order to find the dominant sequence of acoustic cues. As a result, we found that anger has a close relation with intensity level and 2nd formant bandwidth range; joy has a relative relation with the position of 2nd

•제1저자 : 이상민 •교신저자 : 이호준
•투고일 : 2014. 9. 2, 심사일 : 2014. 10. 6, 게재확정일 : 2014. 10. 14.
* 가톨릭대학교 ELP학부대학 창의교육센터(Ethical Leader Path College, Catholic University of Korea)
** 영동대학교 스마트IT학과(Dept. of Smart IT, Youngdong University)

and 3rd formant values and intensity level; sadness has a strong relation only with prosody cues such as intensity level and pitch level; and fear has a relation with pitch level and 2nd formant value with its bandwidth range. These findings can be used as the guideline for find-tuning an emotional spoken language generation system, because these distinct sequences of acoustic cues reveal the subtle characteristics of each emotional state.

▸ Keywords : Emotional Acoustic Cues, Emotional Prosody Structure, Emotional Voice Quality Parameter, Korean Emotional Speech Analysis, Emotional Speech Recognition, Emotional Speech Generation

I. 서 론

During an everyday conversation, we usually choose a specific word or a phrase in order to deliver an intended meaning considering its inherent word sense and mood. But depending on the context information and its pronunciation, inherent word sense can be modified[1].

For example, a Korean word '예' which corresponds to 'yes' in English has an affirmative word sense, whereas '아니요' which means 'no' has a meaning of denial. But in case of answers for negative interrogative sentences, '예' delivers a meaning of denial, and '아니요' an affirmative word sense. In addition, same piece of speech will imply different discourse-pragmatic word senses if it pronounced in different ways. If a word is spoken in an emotional state of anger, then the inherent word sense will become more negative, and if it is pronounced in an emotions state of joy, the meaning will be more positive.

Gravano et al.[2] examined the effect of contextual and acoustic cues in the disambiguation of three discourse functions of the word 'okay'. Seen from the acoustic point of view, this work was only focused on the prosodic structures of the discourse marker 'okay' such as pitch, intensity, duration. But in order to find the relation between different pronunciations and their effects on the modification of word senses, prosody structures and voice quality information such as formants and their bandwidths should be considered in a coordinated manner. One reason is that, for the case of anger and joy, it is wildly believed that they have similar prosody structures, such as fast speech rate, high average pitch value, wide pitch range, and high intensity level, even though they have opposite emotional meanings and distinct pronunciations [3][4][5].

In this paper, we examine the role of the emotional acoustic cues including both prosody and voice quality parameters for the modification of word sense. For this purpose, pitch and intensity values are considered as prosody cues, three formant values and their bandwidth ranges are regarded as voice quality cues [6].

II. Corpus

As an experimental data, we used the Korean emotional speech corpus that is distributed by the Speech Information Technology & Industry Promotion Center (Si-TEC, http://www.sitec.or.kr). This speech corpus was recorded by three professional actors and three actresses in a sound-proof room, and is composed of emotionally neutral ten sentences with six different emotions

(joy, anger, sadness, fear, boredom, and neutral). Detailed information for each speakers are listed in Table 1.

TABLE 1.
Speaker Information

Name	Age	Gender	Acting Experience
CWJ	29	N	7 years
KKS	27	F	8 years
LHJ	24	F	3 years
MYS	25	F	6 years
PYH	28	M	10 years
YJW	31	M	10 years

An AKG C414-B ULS microphone was used with 16KHz sample rate, and each speech was stored as a 16 bit wave file format. The Korean emotional speech corpus had been evaluated by twenty subjects (eighteen males, two females), and anger turned out to be the most perceivable emotion, and fear, the most confusing one. But the overall acceptance rate is more than 80% (average 85.9%) as shown in Table 2.

TABLE 2.
Perception Test Result Done by Twenty Subjects

	Joy	Anger	Sadness	Fear
CWJ	93.5	88.5	85.5	59.0
KKS	90.5	92.0	80.5	85.5
LHJ	67.5	98.0	84.5	88.5
MYS	91.5	90.0	89.5	93.5
PYH	95.0	99.0	94.0	61.5
YSW	89.5	98.5	89.5	93.5
AVG	87.9	94.3	87.3	80.3

In this paper, we used two sentences, '예' and '아니요' which have positive and negative inherent meanings respectively, spoken by six speakers considering five emotions (anger, fear, joy, neutral, and sadness). For each type of sentence, thirty pieces of speech data are provided for data analysis.

III. Data Analysis

For the analysis of the prosody structure of sixty pieces of speech data, we used mean value of each pitch (f0) contour and intensity contour. Before the extraction of f0 contour, we applied a pitch contour smoothing algorithm to reduce noise values and errors.

And for the analysis of voice quality parameters, we extracted continuous 1st, 2nd, and 3rd formant (F1, F2, and F3) values and their band-width (B1, B2, and B3) ranges. For the reason that the analyzed locations or values of the formant are very sensitive according to the predefined environmental parameters, we manually adjusted maximum range of formant (from 4,000Hz to 5,500Hz) and number of formants (4~6) considering their spectrograms. Instead of the use of traditional mean values, we provided mid-hinge values for the three formant locations and their bandwidth ranges in order to minimize analytic errors. The mid-hinge value can be calculated by the average of the first quantile (25%) and the third quantile (75%).

For the accurate and reliable comparison of the acoustic features extracted from different speakers, speaker dependent tendencies should be removed. In this paper, we proposed simple value comparison method for this parameter normalization step. First we set the acoustic features extracted from every speaker in the emotional state of neutral as baseline values. And then we compared the differences between baseline values and corresponding acoustic features of different four emotions. Following Table 3 and Table 4 show the extracted two prosody related features and six voice quality related features for Korean sentences "예" and "아니오" respectively. The "file name" column both in Table 3 and Table 4 consists of one character indicates emotional status such as a: anger, f: fear, h:

happiness/joy, n:neutrality, s:sadness, and two digits implies sentence number, and speakers' ID.

TABLE 3.
Analysis Result of "예"

file name	F0	Int.	F1	F2	F3	B1	B2	B3
a05 cwj	159	74	509	1815	2495	296	147	213
a05 kks	211	68	638	2176	2843	102	129	172
a05l hj	230	75	592	2418	3243	183	223	400
a05 mys	210	63	441	2518	3207	209	178	627
a05 pyh	200	79	525	1827	2464	353	118	194
a05 ysw	292	72	515	1795	2696	139	80	404
f05 cwj	240	73	362	1752	2572	200	278	286
f05 kks	285	74	526	2451	3177	121	227	421
f05 lhj	253	64	371	2431	3006	115	437	357
f05 mys	234	62	377	2661	3258	246	239	583
f05 pyh	196	63	414	1916	2463	533	228	153
f05 ysw	118	60	828	1911	2753	494	231	187
h05 cwj	149	69	419	1990	2689	166	110	161
h05 kks	410	72	868	2317	3098	303	176	345
h05 lhj	202	68	515	2769	3477	132	142	257
h05 mys	244	74	547	2567	3157	155	165	282
h05 pyh	213	67	445	1849	2653	234	161	344
h05 ysw	274	74	531	1855	2723	292	137	286
n05 cwj	94	59	332	2056	2791	73	137	206
n05 kks	149	53	454	2520	2982	95	449	401
n05 lhj	179	56	402	2841	3505	129	276	586
n05 mys	187	61	458	2700	3199	87	378	1204
n05 pyh	96	54	434	2181	3252	337	196	752
n05 ysw	91	53	428	1935	2775	124	168	237
s05 cwj	92	55	562	1923	2621	586	161	300
s05 kks	224	60	448	2200	2849	165	222	163
s05 lhj	283	59	736	2250	2875	399	168	308
s05 mys	150	54	395	2376	2946	252	273	240
s05 pyh	199	59	443	1870	2675	407	262	288
s05 ysw	99	51	346	1900	2672	260	482	371

TABLE 4.
Analysis Result of "아니요"

file name	F0	Int.	F1	F2	F3	B1	B2	B3
a06 cwj	211	75	673	1360	2530	232	233	162
a06 kks	258	71	642	1531	2921	135	179	234
a06l hj	368	82	784	1594	2840	131	204	496
a06 mys	213	65	595	2037	2971	460	366	499
a06 pyh	298	78	609	1371	2370	384	152	205
a06 ysw	281	74	705	1537	2788	165	271	341
f06 cwj	170	65	553	1604	2846	207	321	242
f06 kks	321	71	769	1511	2726	144	244	912
f06 lhj	358	65	659	1634	2697	204	476	327
f06 mys	256	59	700	1976	2861	449	621	237
f06 pyh	238	66	597	1781	2716	242	349	385
f06 ysw	217	69	326	1106	2184	201	501	570
h06 cwj	129	66	598	1498	2527	303	412	287
h06 kks	300	71	670	1382	2714	129	172	713
h06 lhj	287	68	593	1586	2619	150	413	361
h06 mys	233	66	591	1530	2555	146	618	602
h06 pyh	263	59	453	1282	2312	131	216	390
h06 ysw	104	57	608	1460	2623	280	245	385
n06 cwj	99	63	435	1582	2501	219	405	268
n06 kks	148	51	517	1458	2915	208	238	240
n06 lhj	180	60	453	1323	2536	281	535	286

n06 mys	180	55	422	1140	2288	159	761	341
n06 pyh	100	56	397	1670	2379	232	160	214
n06 ysw	92	56	633	1535	2542	329	255	199
s06 cwj	104	56	449	1738	2543	324	479	328
s06 kks	228	57	519	1498	2808	119	264	359
s06 lhj	234	58	569	1525	2680	196	352	178
s06 mys	153	52	468	1439	2535	149	684	487
s06 pyh	159	64	416	1562	2495	299	437	230
s06 ysw	131	54	233	1387	2469	239	229	277

In order to identify dominant emotional acoustic cues for each emotion, we performed discriminant analyses for each group of '예' and '아니요'. Eight acoustic cures such as f0, intensity, F1, F2, F3, B1, B2, and B3 were used as independent variables and four types of emotion were used as grouping variables for discriminant analyses.

Praat (Boersma and Weenink, 2005) was used to extract two prosody cues and six voice quality cues from sixty pieces of speech, and SPSS 16.0 was used for the statistical analyses.

1. Discriminant Analysis for '예'

Generally, for the discriminant analysis, only two types of grouping variables should be appeared in the data. But in this paper, we had four types of grouping variables, such as anger, fear, joy, and sadness. Therefore, we divided discriminant analyses for '예' and '아니요' into four sub-analyses steps respectively. Consequently, for the discriminant analysis for '예' we performed four sub-analyses, like anger vs others, fear vs others, joy vs others, and sadness vs others.

Tables in Figure 1 indicate discriminant analysis results of the emotional state of anger. The canonical correlation value in the first table indicates the overall performance of the acquired discriminant function. And second table shows the classification results done by acquired discriminant function. In this case, the function achieved 66.7% of classification acceptance rate. Third table shows the dominant emotional acoustic cues sorted in order. From the analysis result, we noticed that f0, which is well known for playing an important role for emotional speech generation, is not useful for the discrimination of anger from other emotions. On the other hand, we found that intensity, 2nd bandwidth, and 1st bandwidth information are the dominant acoustic cues for the emotional state of anger.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.650[a]	100.0	100.0	.628

a. First 1 canonical discriminant functions were used in the analysis.

Classification Results

		Emotion	Predicted Group Membership		
			Anger	Others	Total
Original	Count	Anger	4	2	6
		Others	3	15	18
	%	Anger	66.7	33.3	100.0
		Others	16.7	83.3	100.0

Structure Matrix

	Function
	1
intensity	-.515
B2	.311
B1	.285
F2	.280
F1	-.115
F3	.090
B3	-.066
f0	-.018

Fig. 1 Analysis results of anger

Canonical correlation value for fear is 0.438, joy is 0.671, and sadness is 0.777. Following tables in Figure 2 indicate the dominant emotional acoustic cues for fear, joy and sadness.

From these results, we can conclude that each emotion has distinctive order of dominant emotional acoustic cues.

Classification rate for fear is 66.7%, joy is 83.3%

and sadness is 100%.

Fear

	Function 1
B2	.556
F2	.324
F1	-.306
B1	.188
f0	.101
B3	.095
F3	.055
intensity	-.023

Joy

	Function 1
intensity	.354
F2	.353
f0	.333
B2	-.265
F3	.258
B1	-.255
F1	.186
B3	-.054

Sadness

	Function 1
intensity	.745
B1	-.310
f0	.305
F2	.200
B2	-.178
F3	.150
F1	.090
B3	.041

Fig. 2 Structure matrix for three emotions

2. Discriminant Analysis for '아니오'

Following tables in Figure 3 indicate the sequence of dominant acoustic cues for anger, fear, joy, and sadness in order.

Similar to the result of discriminant analysis for '예', each emotion has distinct sequence of acoustic cues. But we can find a weak tendency between analysis results for '예' and '아니요'. For the emotional state of anger, first two dominant acoustic cues for '예' and '아니요' are same as intensity and 2nd bandwidth. And for sadness, intensity and f0 appeared in the top of the list. But for the emotional state of fear, we can only find 2nd bandwidth information in common. Numerical analysis of this tendency will be discussed in Section 4.

Canonical Correlation value for anger is 0.885, fear is 0.659, joy is 0.501, and sadness is 0.715. Also classification rate for anger is 100%, fear is 77.8%, joy is 61.1%, and sadness is 77.8%.

Anger

	Function 1
intensity	-.407
B2	.262
f0	-.212
F1	-.185
F3	-.131
B3	.105
B1	-.063
F2	-.032

Fear

	Function 1
f0	.322
B2	.302
B3	.232
F2	.132
F1	.108
F3	.093
B1	.083
intensity	.039

Joy

	Function 1
B3	-.415
F3	.349
B1	.298
F2	.262
f0	.162
intensity	.104
F1	-.069
B2	.037

Sadness

	Function 1
intensity	.709
f0	.649
F1	.515
B3	.239
B2	-.213
F3	.123
F2	.025
B1	.022

Fig. 3 Analysis result of structure matrix

IV. Discussion

For the numerical representation of the relation we roughly found in Section 3, we assigned accumulative values for every sequence of acoustic cue. For example, acoustic cue 'intensity' in Figure 1 will ranked as 8, 'B2' as 7, 'B1' as 6, and so on. Every sequence of acoustic cue shown in Figure 2 and Figure 3 also marked with appropriate values. Table 5 shows the result.

TABLE 5.
Numerical Representation of Dominant Acoustic Cues

Anger		Fear	
Intensity	16	B2	15
B2	14	f0	12
F1	9	F2	12
B1	8	F1	10
f0	7	B3	9
F3	7	B1	7
F2	6	F3	5
B3	5	Intensity	2

Joy	
F2	12
Intensity	11
F3	11
f0	10
B1	9
B3	9
B2	6
F1	4

Sadness	
Intensity	16
f0	13
F1	8
B2	8
F2	7
B1	7
F3	6
B3	6

As we discussed in Section 3, the emotional states of anger, fear, and sadness formulate their distinct sequence of acoustic cues. And with the help of numerical representation, now we can estimate distinct sequence of acoustic cues for joy as well. Because the maximum value for each case is 8 and the minimum value is 1, we can regard acoustic cue whose numeric representation is more than 12 (third qunatile, 75%) as a dominant acoustic cue.

Based on the observation of these analyzed results, we can conclude that anger has a close relation with intensity level and 2nd bandwidth range, whereas joy has a relation with the position of 2nd and 3rd formant values and intensity level. Also sadness has a strong relation with prosody cues (intensity and f0) rather than voice quality cues, but fear has a relation with pitch level and 2nd formant value with its bandwidth range.

V. Conclusion

In this paper, we examined the role of emotional acoustic cues including both prosody and voice quality parameters for the modification of a word sense. For the extraction of prosody parameters and voice quality parameters, we used 60 pieces of speech data spoken by six speakers with five different emotional states. We analyzed eight different emotional acoustic cues, and used a discriminant analysis technique in order to find the dominant sequence of acoustic cues according to the different emotional states.

As a result, we found distinct sequences of acoustic cues for four different emotional states. These findings can be used as the guideline for fine-tuning an emotional spoken language generation system, because these distinct sequences of acoustic cues reveal the subtle characteristics of each emotional state.

Further analyses of emotional speech data are necessary, taking into account various speakers, speaking environment, and speaking styles. And more organized evaluation and interpretation strategies are essentially needed for further work.

VI. Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning(grant number: NRF-2012R1A1A1013389).

참고문헌

[1] Gi-Jeong Lim, Jung-Chul Lee. 2012. Improvement of Naturalness for a HMM-based Korean TTS using the prosodic boundary information. Journal of The Korea Society of Computer and Information, vol. 17, no. 9, pp. 75-84, September 2012.

[2] Agustín Gravano, Stefan Benus, Héctor Chávez, Julia Hirschberg, and Lauren Wilcox. 2007. On the role of context and prosody in the interpretation of 'okay'. 45th Conference of the ACL, pages 800-807.

[3] Elissaveta Abadjieva, Iain R. Murray, John L. Arnott. 1993. Applying Analysis of Human Emotion Speech to Enhance Synthesis Speech. Eurospeech 93, pages 909-912.

[4] Marc Schröder. 2001. Emotional Speech Synthesis: A Review. Eurospeech 2001, pages 561-564.

[5] Ho-Joon Lee and Jong C. Park. 2009.

Interpretation of User Evaluation for Emotional Speech Synthesis System. Human Computer Interaction International 2009.

〔6〕 Mark Tatham and Katherine Morton. 2006. Expression in Speech: Analysis and Synthesis. Oxford University Press.

저 자 소 개

이 상 민

1998: 가톨릭대학교
국어국문학과 문학사.
2000: 가톨릭대학교
국어국문학과 문학석사.
2004: 가톨릭대학교
국어국문학과 문학박사
현 재: 가톨릭대학교
ELP학부대학 창의교육센터
초빙 교수
관심분야: 글쓰기, 스토리텔링
Email : lsm75@hanmail.net

이 호 준

2001: 한국과학기술원
전산학과 공학사.
2003: 한국과학기술원
전산학과 공학석사.
2010: 한국과학기술원
전산학과 공학박사
현 재: 영동대학교
마트IT학과 조교수
관심분야: 자연언어처리,
감정 음성 합성, 스토리텔링
Email : hjlee@webmail.yd.ac.kr