한글의 초성, 중성, 종성 단위의 조건적 발생확률과 엔트로피

이 재홍 오 상현°
서울대학교 전자공학과

On the Conditional Probabilities and the Entropies of 'Choseong', 'Jungseong', and 'Jongseong' of a Korean Syllable

Jae Hong Lee Sang Hyun Oh

Department of Electronics Engineering
Seoul National University

Abstract

We regard 'choseong', 'jungseong', and 'jongseong' of a Korean syllable as random variables. We calculate the conditional probabilities and the entropies of three units. We obtain mutual information between units.

1. 서론

언어는 인간의 의사소통의 가장 중요한 수단으로서 언어가 문자화된 것이 글이다. 글은 기호들이 연속적으로 나열되어 형성된 기호열로 볼 수 있다. 이렇게 형성된 기호열은 정보를 내포하며 그 척도로서 엔트로피가 사용된다 [1], [2].

Shannon의 정보이론에 근거한 언어의 엔트로피가 몇몇 언어에 대하여 연구되었다 [3]. 한글에 있어서는 음절 단위와 자소 단위의 엔트로피에 관한 연구가 있었다 [4], [5], [6]. 또 엔트로피 연구의 자료가 되는 한글의 음절 및 낱말 단위의 빈도 분포가 조사된 바 있다 [7], [8].

한글은 다른 언어와는 달리 초성, 중성, 종성의 각 자소가 모여서 한 음절을 구성하고 이 음절이 모여서 낱말을 구성하는 특성을 가지고 있다. 이러한 특성에 비추어 볼 때 초성, 중성, 종성을 단위로 하여 엔트로피를 계산하는 것은 의미있는 일이다.

이 논문에서는 한글의 음절별 발생빈도 통계로부터 초성, 중성, 종성 단위의 조건적 발생확률 (conditional probability)을 구하고 그것을 사용하여 초성, 중성, 종성 단위의 조건적 엔트로피 (conditional entropy)를 구한다. 또한 초성, 중성, 종성 단위 간의 평균 상호정보량 (average mutual information)을 계산하고 비교한다.

2. 한글의 엔트로피와 정보량

한 음절을 이루는 초성, 중성, 종성은 각각 그 자소들의 집합에서 어떤 확률을 가지고 선택되는 불규칙 변수(random variable)로 볼 수 있다. 초성, 중성, 종성을 표시하는 불규칙 변수를 각각 X, Y, Z라 하고 그 표본공간(sample space) 즉 자소들의 집합을 각각 Ax, Ay, Az라 하자. 그러면 Ax, Ay, Az는 다음과 같이 주어진다.

Ax = {ㄱ, ㄲ, ㄴ, ㄷ, ㄸ, ㄹ, ㅁ, ㅂ, ㅃ, ㅅ, ㅆ, ㅇ, ㅈ, ㅉ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ}

Ay = {ㅏ, ㅐ, ㅑ, ㅒ, ㅓ, ㅔ, ㅕ, ㅖ, ㅗ, ㅘ, ㅙ, ㅚ, ㅛ, ㅜ, ㅝ, ㅞ, ㅟ, ㅠ, ㅡ, ㅢ, ㅣ}

Az = {공백소, ㄱ, ㄲ, ㄳ, ㄴ, ㄵ, ㄶ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅁ, ㅂ, ㅄ, ㅅ, ㅆ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ}

Ax ∩ Az = {ㄱ, ㄲ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅆ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ}

종성에는 공백소가 발생할 수 있는데, 공백소는 종성에 자음이 없는 경우를 표시한다. 그러나 초성에는 공백소가 발생하지 않는다.

한 음절 내의 단위 간의 발생의 상관관계를

알아보기 위하여 먼저 조건적 발생확률과 조건적 엔트로피를 구한다. 초성 X가 주어질 때 중성 Y의 조건적 엔트로피 H(Y|X)는 다음 식으로 주어진다.

$$H(Y|X) = -\sum_{x} p(x)\, H(Y|X=x)$$
$$= -\sum_{x} p(x) \sum_{y} p(y|x) \log p(y|x) \quad (1)$$

초성의 엔트로피 H(X)는 다음 식으로 주어진다.

$$H(X) = -\sum_{x} p(x) \log p(x) \quad (2)$$

식 (1)과 (2)로부터 초성과 중성 간의 평균 상호정보량 I(X;Y)는 다음 식으로 주어진다.

$$I(X;Y) = H(X) - H(X|Y) \quad (3)$$

같은 방법으로 중성과 종성, 초성과 종성간의 평균 상호정보량이 구해진다.

3. 계산 및 결과

초성, 중성, 종성 간의 조건적 발생확률을 계산하는데 사용된 자료는 '우리말 역순 사전'에 조사된 전체 음절별 누적 빈도수를 자료로 사용하였다[8]. 음절별 빈도수로부터 p(x,y,z)가 계산되고, p(x,y,z) 로부터 p(y|x)가 다음 식에 의하여 구해진다 [9],[10].

$$p(y\,x) = \frac{p(x,y)}{p(x)} = \frac{\sum_{z} p(x,y,z)}{\sum_{y,z} p(x,y,z)} \quad (4)$$

계산된 p(y|x)은 표 1과 같다. 같은 방법으로 계산된 p(x|y), p(z|y), p(y|z), p(z|x), p(x|z)는 각각 표 2에서 6까지와 같다.

초성이 특정한 값 x로 주어졌을 때의 중성의 엔트로피 H(Y|X=x)와 이것을 x에 대하여 평균한 조건적 엔트로피 H(Y|X)는 식 (1)로부터 구해지고 계산 결과는 표 7과 같다. 표 7에서 중성의 엔트로피는 초성이 'ㅇ'일 때 가장 크고 초성이 'ㅎ'일 때 가장 작음을 볼 수 있다. 같은 방법으로 계산된 H(X|Y), H(Z|Y), H(Y|Z), H(Z X), H(X|Z)는 각각 표 8에서 12까지와 같다.

초성의 엔트로피 H(X)는 식 (2)로부터 구해지고 H(Y), H(Z)도 같은 방법으로 구해지며 계산 결과는 표 13과 같다. 표 13에서 초성이 그 발생의 불확실성이 가장 크고 종성의 불확실성이 가장 작음을 볼 수 있다.

표 13. 각 단위의 엔트로피
table 13. entropy of the units

엔트로피	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

표 14. 평균 상호정보량
table 14. mutual information

상호 정보량	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]

초성과 중성 간의 평균 상호정보량 I(X;Y)는 식 (3)으로부터 구해지고 I(Y;Z), I(X;Z)도 같은 방법으로 구해지며 계산 결과는 표 14와 같다. 표 14에서 평균 상호정보량은 초성과 중성 간이 가장 크고 초성과 종성 간이 가장 작음을 볼 수 있다. 즉 초성이 중성에 대하여 공급하는 정보량이 가장 크고 초성이 종성에 대하여 공급하는 정보량이 가장 작음을 볼 수 있다.

4. 결론

한글의 음절을 초성, 중성, 종성의 단위로 나누고, 이들을 각각 불규칙 변수로 간주하였다. 각 단위의 조건적 발생확률을 계산하고, 이것을 사용하여 조건적 엔트로피를 계산하였다. 또 각 단위의 엔트로피를 계산하여, 초성의 엔트로피가 가장 크고 중성의 엔트로피가 가장 작음을 보였다. 끝으로 각 단위들 간의 평균 상호정보량을 비교하여 초성과 중성의 상호 정보량이 가장 큼을 보였다.

인용 문헌

[1] C. Shannon, "A mathmatical theory of communication," Bell Syst. Tech. J., vol. 27, pp. 379-413, 623-656, July and Oct. 1948.

[2] C. Shannon, "Prediction and entropy of printed English," Bell Syst. Tech. J., vol. 30, pp. 50-64, Jan. 1951.

[3] R. Manfrino, "Printed protuguese entropy-statistical calculation," IEEE Trans. Inform. Theory, vol. IT-16, p. 122, Jan. 1970.

[4] 이 주근, 최 흥문, "한국어 음절의 entropy에 관한 연구," 전자공학회지, 제11권, 제3호, pp. 15-21, 1974년 6월

[5] 안 수길, 안 지환, "공백소를 포함한 한글 자소발생 확률과 엔트로피," 전자공학회지, 제17권, 제2호, pp. 23-28, 1980년 4월

[6] 남궁 건, 한글 낱말의 발생빈도 분포의 엔트로피에 관한 연구. 석사 학위 논문, 서울대학교, 1979.

[7] 문교부, 우리말 말수 사용의 잦기 조사. 문교부 1956.

[8] 유 재원, 우리말 역순사전. 정음사 pp. 858-877, 1985

[9] R. McEliece, The Theory of Information and Coding. Addison-Wesley, 1977.

[10] R. Gallager, Information Theory and Reliable Communication. Wiley, 1968.

표

표 2. 조건적 발생확률 p(x|y)
table 2. conditional probability p(x|y)

x \ y	ㅏ	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ	ㅚ	ㅛ	ㅜ	ㅝ	ㅞ	ㅟ	ㅠ	ㅡ	ㅢ	ㅣ
ㄱ	.18839	.04495	.00291	.00027	.20636	.01307	.03027	.00521	.00324	.01259	.00305	.00352	.00066	.00477	.00068	.00027	.02086	.00614	.14416	—	.14640
ㄲ	.18129	.05287	.00752	—	.11276	.01530	.00380	—	.19483	.00019	.00064	.01225	—	.15790	.00501	.00013	.00362	—	.16631	—	.00182
ㄴ	.36621	.00392	.01614	—	.07748	.01512	.02249	.00019	.11413	—	—	.00227	.00087	.00486	.00019	.00019	.00586	.00170	.00616	.00472	.00579
ㄷ	.21019	.00036	—	—	.11717	.02178	—	—	.18304	.00032	.00154	.02017	—	.11588	.00021	.00054	.03480	—	.18538	—	.01809
ㄸ	.22205	.00286	—	—	.17784	.04124	—	—	.00239	.00150	.00079	.00278	—	.13204	—	—	.02536	—	.19586	.00436	.02181
ㄹ	.12521	.03482	.00435	—	.12369	.04006	.02564	.00322	.04586	.00017	—	.00113	.00130	.05254	—	—	.00009	.00096	.13621	—	.46383
ㅁ	.20467	.00430	.03314	—	.11242	.02186	.02090	—	.00662	—	—	.00066	.00027	.28410	.00053	—	.00013	—	.00278	—	.00596
ㅂ	.31442	.00013	.00286	—	.12932	.00042	.04279	—	.10477	.00043	—	.00143	—	.29906	.00012	—	—	—	.00072	—	.11514
ㅃ	.23216	.11178	.01376		.17283		.00082	—	.05345	—	—	—	.01092	.16879	—	—	—	.01032	.01290	—	.11846
ㅅ	.21387	.00242	—	—	.12635	.03806	.00033	—	.12821	.00055	.00055	.02335	.00033	.11553	.00011	—	.00502	—	.00086	—	.16367
ㅆ	.31824	.03813	—	—	.13741	.00131	—	—	.05381	.00197	.00592	.00329	—	.00941	—	.00253	—	—	.11045	.00685	.23077
ㅇ	.13216	.00055	.03828	.00052	.13381	.01184	.04559	.00444	.08764	.01516	.00481	.01443	.01525	.00977	.01362	.00287	.01821	.00770	.07004	.00474	.20174
ㅈ	.25104	.01053	—	—	.15366	.01079	.00044	—	.06115	.00184	.00044	.00327	—	.12600	—	.00035	.01016	—	.00586	—	.34674
ㅉ	.20069	.03044	—	—	.14675	.00717	—	.00066	.12366	.00475	—	.00290	—	.14605	—	—	—	—	.00777	—	.22222
ㅊ	.19536	.06083	—	—	.18143	.01383	.00619	—	.06131	—	—	.00082	—	.11086	—	—	.00342	—	.01146	—	.35288
ㅋ	.17047	.02853	.01587	—	.11246	.00778	.03464	.00173	.21006	.01036	.00433	.00173	—	.05277	.00952	.00067	.03281	—	.21453	—	.07743
ㅌ	.23249	.05431	—	—	.17294	.01818	—	—	.21010	—	.00005	.01236	—	.12292	—	.00005	.01143	—	.13346	.00005	.02946
ㅍ	.27947	.05430	.00132	—	.07947	—	.00402	.00177	.07417	—	—	—	.02340	.30617	—	—	—	—	.00795	—	.10798
ㅎ	.64618	.03380	.00366	—	.04313	.01079	.00625	.00027	.05247	.03090	.00320	.01052	—	.02304	.00146	.00048	.01185	.00226	.03640	.01425	.00432
P(y)	.24250	.00201	.00910	.00000	.13307	.01873	.02001	.00156	.00484	.00060	.00170	.00762	.00264	.12627	.00208	.00063	.01112	.00131	.00705	.00188	.10475

표 1. 조건적 발생확률 p(y|x)
table 1. conditional probability p(y|x)

y \ x	ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ	ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ㅏ	.00312	.02276	.00775	.06640	.01050	.05076	.07037	.00222	.00944	.05453	.01681	.05243	.00031	.01748	.02503	.00713	.01700	.02213	.10061
ㅐ	.10823	.03040	.00101	.15004	.03374	.06487	.07006	.00454	.02119	.13006	.00045	.02163	.03642	.00904	.03506	.00538	.01030	.00005	.04130
ㅑ	.04067	.02516	.00047	—	—	.04600	.23206	.02237	.01401	—	—	.46163	—	—	—	.01678	—	.00220	.02703
ㅒ	.38384	—	—	—	—	—	—	—	—	—	—	.53636	—	—	—	—	—	—	—
ㅓ	.18001	.02500	.02612	.00936	.02454	.00052	.05482	.00012	.01200	.07377	.01331	.11402	.11070	.01546	.03535	.00020	.02312	.01147	.02004
ㅔ	.08737	.02535	.03622	.00190	.04700	.21277	.07100	.03676	—	.12721	.00001	.07243	.10140	.00543	.02026	.00407	.01720	—	.03607
ㅕ	.16119	.00587	.04024	—	—	.11950	.00437	.14532	.04621	.00122	—	.24970	.00203	—	.01216	.01021	—	.05475	.02513
ㅖ	.41640	—	.00543	—	—	.28107	—	—	—	—	—	.32609	—	.00543	—	.01087	—	.02174	.01067
ㅗ	.12308	.00354	.05306	.13043	.02003	.04026	.03006	.07053	.00345	.10400	.00733	.00106	.05182	.01049	.02002	.02284	.03040	.01301	.03520
ㅘ	.23095	.04500	—	.00410	.00557	.00270	—	.00075	—	.00000	.00418	.28412	.02646	.01114	—	.01671	—	—	.32312
ㅙ	.22509	.12000	—	.07000	.01000	—	—	—	—	.02500	.04500	.32500	.02500	—	—	.02500	.01000	—	.12000
ㅚ	.05704	.04004	.01335	.20012	.00779	.01440	.00550	.01335	—	.23404	.00550	.21001	.04116	.00336	.00334	.00222	.02002	—	.00700
ㅛ	.03185	—	.00055	—	—	.04777	.00637	—	.07000	.00055	—	.05005	—	—	—	—	—	.10479	—
ㅜ	.00049	.03906	.03534	.07612	.02347	.04257	.15104	.12363	.01316	.07464	.00050	.00549	.10058	.01734	.03073	.00430	.01618	.04020	.01210
ㅝ	.05235	.07317	.00487	.00013	—	—	.01020	.00407	—	.00407	—	.74707	—	—	—	.04472	—	—	.04472
ㅞ	.05405	.28730	.01351	.00757	—	—	—	—	—	—	.05405	.37038	.05405	—	—	.01351	.02703	—	.04054
ㅟ	.23476	.00081	.02363	.24619	.04070	.00076	.00076	—	—	.03506	—	.18750	.00765	—	.01067	.02020	.01020	—	.05704
ㅠ	.01290	—	.05600	—	—	.07007	—	—	.07742	—	—	.67007	—	—	—	—	—	—	.10060
ㅡ	.28732	.05981	.04857	.17733	.04611	.15250	.00204	.00058	.00146	.00432	.01636	.10400	.00623	.00127	.00400	.02415	.02727	.00175	.02044
ㅢ	—	—	.11281	—	.04055	—	—	—	—	—	.05436	.20029	—	—	—	—	.00081	—	.48100
ㅣ	.00023	.01019	.02003	.00612	.00252	.21306	.03322	.04432	.00619	.00402	.01610	.10067	.13404	.01707	.00033	.00413	.00200	.01101	.02216
P(x)	.12519	.03044	.04406	.07900	.02130	.00746	.06304	.07112	.00066	.07761	.01280	.11454	.00593	.01419	.03107	.00900	.01779	.01020	.00365

표 3. 조건적 발생확률 p(z|y)
table 3. conditional probbaility p(z|y)

y \ z	공백소	ㄱ	ㄲ	ㄳ	ㄴ	ㄵ	ㄶ	ㄷ	ㄹ	ㄺ	ㄻ	ㄼ	ㄽ	ㄾ	ㄿ	ㅀ	ㅁ	ㅂ	ㅄ	ㅅ	ㅆ	ㅇ	ㅈ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
ㅏ	.45433	.00009	.00031	.00008	.07418	.00185	.00171	.00002	.13472	.00297	.00009	.00073	—	.00017	—	.00028	.04849	.03462	.00140	.01192	.00017	.11239	.01168	.00136	—	.01059	.00336	.00311
ㅐ	.75884	.04050	—	—	.01005	—	—	—	.01380	—	—	—	—	—	—	—	.02233	.00424	—	.05037	.00000	.00082	—	—	—	.00033	—	—
ㅑ	.24077	.33371	—	—	.01084	—	—	—	.04753	—	—	.00003	—	—	—	—	.02423	.00093	—	.00258	—	.31834	—	—	—	.00746	—	.00408
ㅒ	1.00000	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
ㅓ	.37737	.15034	.00210	.00032	.05018	.00045	—	.00471	.13218	.00217	.00038	.00446	—	—	—	—	.03586	.03784	.01178	.03478	.00019	.11855	.00433	.00013	.00045	.00581	.00358	.00427
ㅔ	.85514	.00615	—	—	.01720	—	—	—	.00272	—	—	—	—	—	—	—	.02540	.00001	—	.04060	.00001	.04029	—	—	—	—	—	—
ㅕ	.32671	.00606	.00305	—	.17833	—	—	.00383	.07004	—	—	.00001	—	—	—	—	.04135	.04662	.00001	.04985	.00122	.15291	—	.00263	.00567	.03640	.01176	.00283
ㅖ	.84783	—	—	—	.00543	—	—	—	—	—	—	—	—	—	—	—	—	.00543	—	.10870	.03261	—	—	—	—	—	—	—
ㅗ	.38583	.03301	.00250	.00030	.05736	—	.00009	.00327	.13002	.00071	.00054	—	.00054	.00018		.00118	.04838	.02296		.02743		.20011	.00313	.02412		.00025	.00277	.00780
ㅘ	.39415	.07342	—	—	.13849	—	—	—	.14007	—	—	—	—	—	—	—	.00136	—	—	.01950	.00270	.23120	—	—	—	—	—	—
ㅙ	.08500	.00000	—	—	.02500	—	—	—	.00500	—	—	—	—	—	—	—	—	—	—	.02000	—	.20300	—	—	—	—	—	—
ㅚ	.07542	.00550	—	—	.05339	—	—	—	.00334	—	—	—	—	—	—	—	.01357	.00222	—	.03071	—	.00779	—	—	—	—	—	—
ㅛ	.08471	.05732	—	—	—	—	—	—	.00037	—	—	—	—	—	—	—	.00637	—	—	.01582	—	.22030	—	—	—	—	—	—
ㅜ	.42071	.11615	.00083	—	.07471	—	—	.00400	.28550	.00251	.00035	—	—	.00049	—	.00148	.03841	.00944	—	.03425	—	.00662	.00190	.00134	—	.00040	.00079	—
ㅝ	.35304	.03252	—	—	.25283	—	—	—	.28062	—	—	—	—	—	—	—	—	—	—	.00487	—	.00584	—	—	.00487	—	—	—
ㅞ	.47297	.24324	—	—	.00757	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	.21622	—	—	—	—	—	—
ㅟ	.00335	.00229	—	—	.00913	—	—	—	.00510	—	—	—	—	—	—	—	.01524	.00762	—	.14839	—	.00686	—	—	—	—	—	—
ㅠ	.43226	.18710	—	—	.03220	—	—	—	.03006	—	—	—	—	—	—	—	—	—	—	—	—	.17418	—	.11613	—	—	—	—
ㅡ	.33006	.04441	—	—	.13205	—	.00138	.00331	.18775	.00023	.00019	—	—	—	.00010	.00040	.18379	.01246	—	.00084	—	.03380	.00390	.00010	.00010	.00028	.00156	.00028
ㅢ	.79738	—	—	—	.18818	—	—	—	—	—	—	—	—	—	—	—	.01351	—	—	—	—	.00001	—	—	—	—	—	—
ㅣ	.72324	.01806	—	—	.03518	—	—	.00050	.09282	.00037	.00000	—	—	—	—	.00023	.04022	.02325	—	.02094	.00004	.01000	.00240	.00252	—	.00362	.00078	.00023
P(y)	.49075	.00143	.00123	.00027	.00536	.00051	.00056	.00378	.12537	.00220	.00032	.00089	.00005	.00012	.00002	.00044	.05073	.02100	.00192	.03144	.00033	.00090	.00473	.00340	.00019	.00500	.00363	.00021

표 4. 조건적 발생확률 p(y|z)
table 4. conditional probability p(y|z)

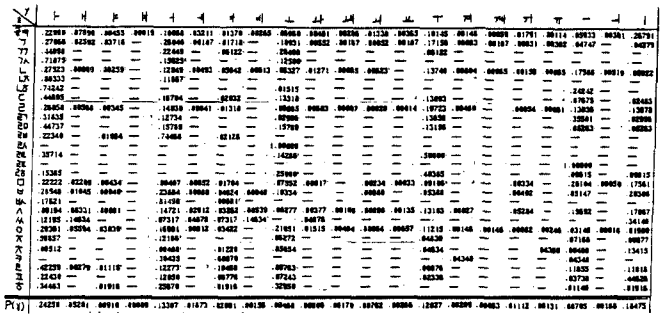

표 5. 조건적 발생확률 p(z|x)
table 5. conditional probability p(z|x)

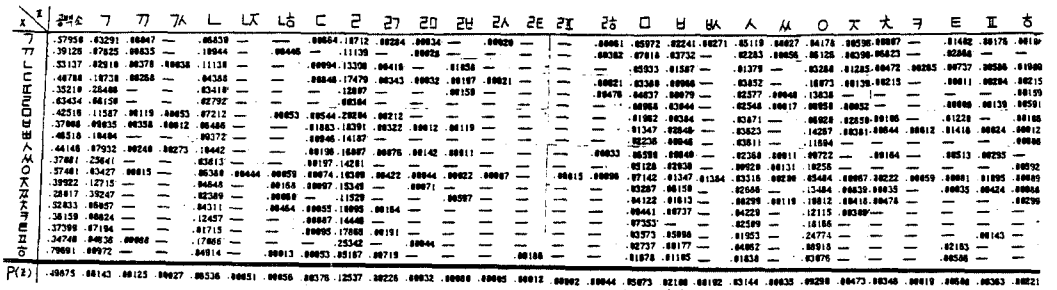

표 6. 조건적 발생확률 p(x|z)
table 6. conditional probability p(x|z)

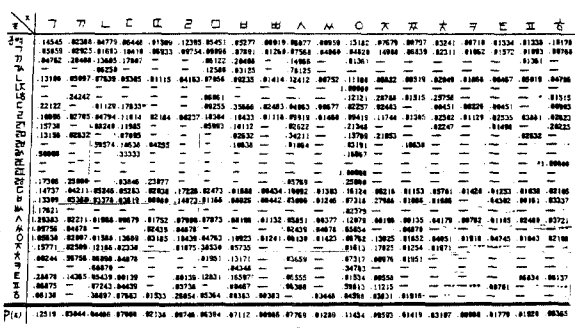

표 7. 조건적 엔트로피 H(Y|X)
table 7. conditional entropy H(Y|X)

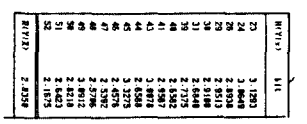

표 8. 조건적 엔트로피 H(X|Y)
table 8. conditional entropy H(X|Y)

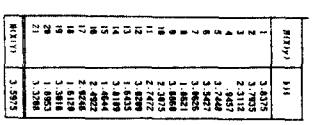

표 9. 조건적 엔트로피 H(Z|Y)
table 9. conditional entropy H(Z|Y)

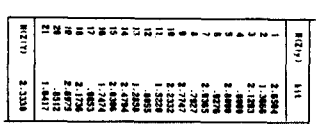

표 10. 조건적 엔트로피 H(Y|Z)
table 10. conditional entropy H(Y|Z)

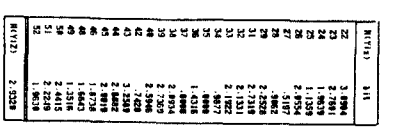

표 11. 조건적 엔트로피 H(Z|X)
table 11. conditional entropy H(Z|X)

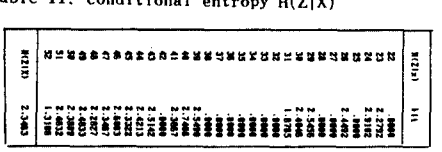

표 12. 조건적 엔트로피 H(X|Z)
table 12. conditional entropy H(X|Z)