말소리와 음성과학 제7권 제1호 (2015.03.31)
ISSN 2005-8063, pp. 97~105

www.speechsciences.or.kr
http://dx.doi.org/10.13064/KSSS.2015.7.1.097

영어 동시발화의 자동 억양궤적 추출을 통한 음향 분석

An acoustical analysis of synchronous English speech using automatic intonation contour extraction

이 서 배[1)]
Yi, So Pae

ABSTRACT

This research mainly focuses on intonational characteristics of synchronous English speech. Intonation contours were extracted from 1,848 utterances produced in two different speaking modes (solo vs. synchronous) by 28 (12 women and 16 men) native speakers of English. Synchronous speech is found to be slower than solo speech. Women are found to speak slower than men. The effect size of speech rate caused by different speaking modes is greater than gender differences. However, there is no interaction between the two factors (speaking modes vs. gender differences) in terms of speech rate. Analysis of pitch point features has it that synchronous speech has smaller Pt (pitch point movement time), Pr (pitch point pitch range), Ps (pitch point slope) and Pd (pitch point distance) than solo speech. There is no interaction between the two factors (speaking modes vs. gender differences) in terms of pitch point features. Analysis of sentence level features reveals that synchronous speech has smaller Sr (sentence level pitch range), Ss (sentence slope), MaxNr (normalized maximum pitch) and MinNr (normalized minimum pitch) but greater Min (minimum pitch) and Sd (sentence duration) than solo speech. It is also shown that the higher the Mid (median pitch), the MaxNr and the MinNr in solo speaking mode, the more they are reduced in synchronous speaking mode. Max, Min and Mid show greater speaker discriminability than other features.

Keywords: synchronous speech, intonation contour, pitch range, pitch slope, pitch distance

1. 서 론

많은 연구자들이 일관성 있고 보편타당한 음성학적 정보를 얻기 위해 개별 화자로부터 말미암은 발화 변이성(variation)의 영향을 어떻게 극복하고 화자공통적인 특성을 이해할 것인가에 관심을 갖고 있다. 화자인식과 같은 음성신호처리 분야에서도 어떻게 하면 화자 내 변이의 부정적 영향을 최소화하면서 화자 간의 변별력을 높일 수 있는가가 중요한 연구이슈이다. 이런 맥락에서 발화들 간의 시간적 변이(temporal variation)를 감소시키는 것으로 알려진 동시발화는 주목할 만하다. 동시발화 연구는 두 화자가 같은 내용의 텍스트(익숙하지 않은)를 함께 동시에 읽었을 때 얻어지는 발화를 분석하는 것으로 Cummins (2000)에 의해 소개되었다.

일상생활에서 예를 들면, 국민의례에서 국기에 대해 맹세할 때, 종교의식에서 회중들이 주기도문, 사도신경 등등을 함께 외울 때와 회중들이 같은 본문(성경, 불경)을 합독할 때 그리고 시위대의 동시발화(미국의 일부 주에서는 시위에 확성기를 사용할 수 없으므로 자신들의 주장을 여러 명이 동시에 종종 외침)등등의 경우에 동시발화를 접할 수 있다. 동시발화를 통해 화자는 자신만의 고유한 발화특성을 최소화하고 서로가 공유하고 있다고 동의하는 보편적 특성이 드러나는 방향으로 자신의 발화를 통제하고 수정할 것으로 예측된다(김미란 & 남호성, 2012). 이러한 동시발화는 영어리듬의 교육에서도 효과를 보이고 있어 동시발화의 교육적 연구 가치 또한 최근 부각되고 있다(Banzina, et al., 2014).

지금까지 동시발화의 리듬과 시간적 변이에 관련한 연구는 비교적 활발히 이루어져 왔다(Cummins, 2000; Cummins & Roy, 2001; Cummins, 2003; Cummins, et al., 2006; 김미란 & 남호성, 2012; Cummins, et al., 2013). 그러나 동시발화의 억

1) 창원대학교, 영어영문학과 sopaeyi@pusan.ac.kr

접수일자: 2015년 1월 31일
수정일자: 2015년 3월 10일
게재결정: 2015년 3월 11일

양에 관한 연구는 동시발화에서 피치변위가 상당수준으로 감소한다는 분석(Cummins, 2000; Cummins & Roy, 2001)이외에는 거의 보고되고 있지 않는 실정이다.

이러한 배경에서 동시발화의 시간적 변이감소가 억양에 어떠한 영향을 미치는지 살펴보고 혼자서 낭독하는 단독발화와 두 명이 함께 낭독하는 동시발화의 억양을 비교분석하는 연구는 의미 있는 일이라 할 수 있다. 이를 위해 본 연구는 먼저 대용량 동시발화 음성코퍼스에서 억양궤적을 자동으로 추출하기 위해 Momel(Hirst, 2000)을 사용하였다. 자동 억양궤적 추출 알고리듬인 Momel은 영어, 프랑스어, 독일어, 스페인어 등의 언어를 대상으로 한 평가에서 높은 성능을 보여 주었다(Campione, 2001). 사람이 주관적 판단으로 억양궤적을 찾는 것에 비해 Momel은 객관성과 일관성을 가지고 억양궤적을 높은 정확도로 추출하고 있으며, F0값 추정이 힘든 무성자음이나 무성화된 모음에서도 억양목표점을 찾을 수 있어서 최근 발화속도와 관련한 억양의 음향특징 연구에도 사용되었다(이서배, 2014b).

<그림 1>은 본 연구에서 연구목적에 맞도록 수정한 Momel의 praat 스크립트(부록 참조)를 사용해 자동으로 추출한 억양궤적을 보여준다. <그림 1>에서 “Play in the street up ahead.”라는 문장의 발화를 praat(Boersma, 2001)의 기본 알고리듬으로 계산한 F0값들이 중간에 나타나 있고 Momel이 추출한 억양궤적이 맨 밑에 있다. 이것은 원 억양궤적(중간그림)에 나타난 수많은 피치값들 중에 대표성을 가지는 몇 개의 값을 추정해 억양궤적을 단순화시킨 것이다. 이 예에서 맨 밑의 6개 피치포인트만 가지고 음성합성으로 원래음성을 복원해도 실제 원래음성과 비교하는 간단한 청취실험 결과 억양의 차이가 느껴지지 않는 것으로 나타났다.

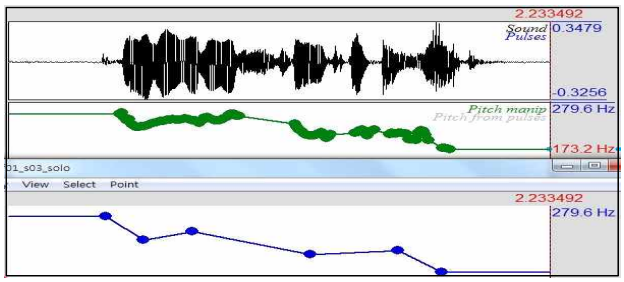

그림 1. Momel을 이용한 자동 억양 궤적추출
Figure 1. Automatic extraction of intonation contour by Momel

2. 실 험

2.1 음성 코퍼스

본 연구는 화자의 특성과 화자인식 성능향상을 연구하기 위해 개발된 CHAINS 음성 코퍼스(Cummins et al., 2006)를 이용하였다. CHAINS 음성 코퍼스는 음향처리된 스튜디오에서 Neumann U87 콘덴서 마이크로 녹음해 16bit 44.1kHz PCM 인코딩된 WAV파일 형태로 저장하였는데 본 연구는 이 음성 코퍼스에서 CSLU's Phonetically Rich Phrases의 9문장과 TIMIT sentences의 24문장을 합한 총 33문장(Cummins et al., 2006 부록 참조)을 28명의 영어 원어민(남성 16명 + 여성 12명)[2]이 단독발화(solo speech)와 동시발화(synchronous speech)의 형태로 발화한 1,848(33x28x2)개의 발화문장을 분석대상으로 삼았다. 단독발화는 혼자 낭독하였고 동시발화는 2명씩 짝을 이루어 진행되었는데 참가자는 두꺼운 유리벽을 사이에 두고 상대 화자의 모습을 보고 상대 화자가 낭독하는 목소리를 들으면서 동일한 텍스트를 상대 화자와 동시에 낭독하였다.

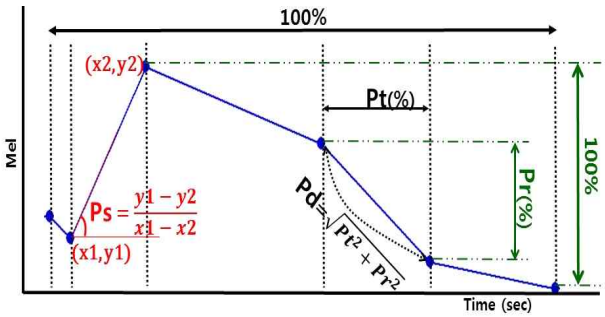

그림 2. 피치포인트 간 억양자질들: 오재혁(2014a)을 재구성
Figure 2. Intonation features between pitch points: adapted from Oh (2014a)

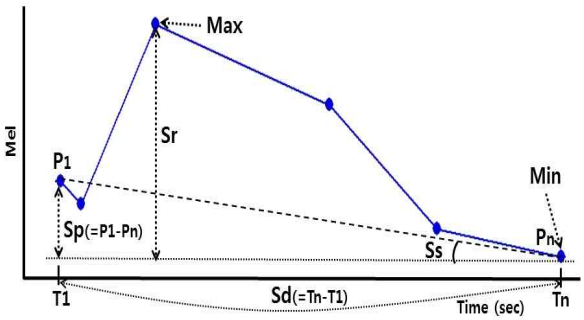

그림 3. 문장수준 억양자질들: 이서배(2014b)를 재구성
Figure 3. Intonation features at sentence level: adapted from Yi (2014b)

2.2 분석 방법

성능 향상된 Momel 버전(Hirst, 2007, Looze & Hirst, 2008)은 praat의 풀다운 메뉴들의 서브메뉴들로 변환되고 이 서브메뉴들이 단계별로 쓰여 지도록 되어 있어서 개별발화 분석에는 편리하지만 대용량발화 분석에는 적합하지 않다. 그러므로 본 연구는 2008년 Momel 저자에 의해 개발된 여러 개의 플러그인(plug-in)용 praat 스크립트들을 본 연구의 특성에 맞도록 통

2) 모두 아일랜드의 더블린과 그 근방에서 Eastern Hiberno-English를 구사하는 화자들로 구성되었다.

합하고 수정하여 하나의 스크립트로 모든 과정을 한 번에 처리할 수 있게 하였다(부록 참조).

동시발화의 억양궤적 추출을 위해 계산한 피치 값은 기존연구(Cummins, 2000; Cummins & Roy, 2001)에서 사용한 Hz가 아니라 청각적 스케일 단위인 mel을 사용하였다[3]. 음성인식, 화자인식 등의 공학시스템에서도 성능향상을 위해 대부분 필터디자인에 Hz대신 mel 스케일을 사용하고 있다. 이러한 점을 감안하면 동시발화의 억양분석에서도 mel을 피치단위로 사용하는 것이 활용범위가 더 클 것이다.

본 연구는 대용량발화에서 억양궤적을 추출하여 억양의 음향특성을 분석한 연구들에서 살펴본 발화문장 내 억양자질들(오재혁, 2014a; 오재혁, 2014b; <그림 2> 참조)과 발화문장 전체에서 구해지는 자질들 즉, 한 문장 당 하나씩 구해지는 억양자질들(이서배, 2014a; 이서배, 2014b; <그림 3> 참조)의 값을 계산하여 억양분석에 사용하였다. 모든 억양자질들의 설명은 아래와 같다.

Pt: 피치포인트[4] 이동시간(피치포인트들 간 이동시간의 백분율)
Pr: 피치포인트 피치변위(피치포인트들 간 피치 차이 절댓값의 백분율)
Ps: 피치포인트들 간 기울기(절댓값)
Pd: 피치포인트들 간의 거리
Sr: 문장 피치변위(문장 내 최대 피치포인트 값과 최소 피치포인트 값의 피치 차이 절댓값)
Sp: 첫째 피치포인트(P1)와 마지막 피치포인트(Pn)의 피치 차이 절댓값
Sd: 첫째 피치포인트에서 마지막 피치포인트까지의 시간 (Tn - T1)
Ss: 문장 피치기울기(Ss = Sp / Sd)
Max: 한 문장 내 최대 피치포인트 값(피치)
Min: 한 문장 내 최소 피치포인트 값(피치)
Mid: 한 문장의 피치포인트 값들 중 중앙값(피치)
MaxNr: Max와 그 문장 중앙값과의 차이(절댓값)
MinNr: Min과 그 문장 중앙값과의 차이(절댓값)

위에서 Max와 Min의 경우, 두 화자 집단(남녀)의 억양 비교 시 남녀의 생리적 차이(성대 크기, 성도 길이, 방패연골 각도)로 인한 F0의 차이를 정규화 할 필요가 있다. 그래서 Max와 Min을 한 문장의 피치 중앙값(Median pitch value)과의 차이로 정규화한 MaxNr과 MinNr을 계산했다. 중앙값은 일반적으로 사용되는 평균값보다 피치의 분석에서와 같이 오류와 정확도의 문제로 인해 종종 생기는 극단적 값들(outliers)의 부정적 영향에 강하므로 정규화가 필요한 억양연구에 쓰여 지고 있다(Forsell, 2007, 이서배 & 김수정, 2011).

3) 변환식 $mel = 1127.01048 \times \ln(1 + f0/700)$을 사용함. 여기서 f0는 기본주파수이고 mel은 청각스케일의 음높이 단위이다.
4) 본 연구에서는 억양목표점(pitch target)이라는 용어를 피치포인트로 대치하였다.

3. 결과 및 분석

3.1 동시발화 유무와 남녀에 따른 발화속도

발화속도는 한 문장의 음절수를 시간으로 나누어 얻은 초당 음절수(syl/sec)로 측정했다. 동시발화 유무(동시발화 vs. 단독발화)와 두 화자 집단(남녀 화자)을 모수 요인으로 하고 초당 음절수를 종속변수로 하는 분산분석(ANOVA)을 시행했는데 동시발화 유무가 통계적으로 유의한 발화속도 차이를 가져오는 것으로 나타났다[$F(1, 171.042)$, $p<0.001$, $\eta_p^2=0.085$](<그림 5> 참조). 그러나 동시발화 유무와 화자 집단 간의 상호작용은 통계적으로 의미가 없었다[$F(1, 2.849)$, $p=0.092$].

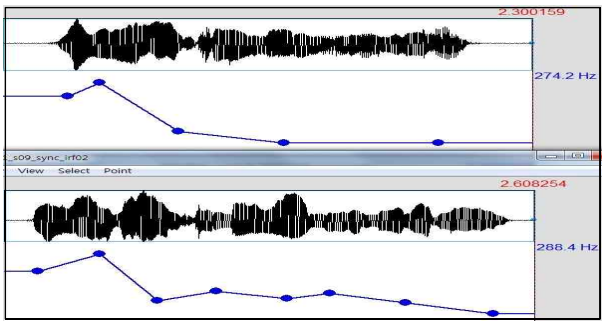

그림 4. 동일 화자가 한 문장(Here I was in Miami and Illinois.)을 단독발화(위)와 동시발화(아래)로 발화한 것의 파형과 피치포인트들

Figure 4. Waveforms and pitch points of solo (top) and synchronous (bottom) utterances of a sentence read by a speaker.

단독발화의 평균은 4.90syl/sec이고 동시발화의 평균은 4.39syl/sec이었다. 기존연구는 발화속도 차이를 느끼게 하는 최소 발화속도 차이(Just Noticeable Difference: JND)를 5%로 보고하고 있다(Quené, 2007). 단독발화와 동시발화의 평균속도 차이는 0.51syl/sec로서 단독발화를 기준으로 5%인 0.24syl/sec보다 큰 차이를 보이고 있으므로 인지적 관점에서도 차이가 난다고 말할 수 있다. 즉, 사람이 들어보아도 단독발화에 비해 동시발화는 확연히 느리다는 것을 인지할 수 있다는 것이다. 기존연구에서는 단독발화보다 동시발화의 속도가 감소하는 현상을 두 화자가 공시화(synchrony) 과제를 수행하면서 상대방과 자신의 발화를 온라인으로 모니터링하며 조정(accommodation)하는 과정 때문에 걸리는 시간이 발화속도 저하를 야기하는

것으로 설명하고 있다(김미란 & 남호성, 2012; Cummins, 2003).

두 화자집단에 따라 발화속도를 분석하면 여성의 평균은 4.51syl/sec, 남성의 평균은 4.78syl/sec로서 두 집단은 통계적으로 유의한 차이가 있는 것으로 나타났다[F(1, 47.270), p<0.001, η_P^2=0.025]. 그러나 효과크기(effect size)로 보면, 동시발화의 유무에 따른 발화속도의 차이(η_P^2=0.085)보다 남녀에 따른 발화속도의 차이(η_P^2=0.025)가 더 작다고 말할 수 있다. 인지적 관점에서도 남녀 간의 차이인 0.27syl/sec는 여성 발화속도를 기준으로 5%인 0.23syl/sec와 남성 발화속도를 기준으로 5%인 0.24syl/sec와 비교해서 조금 차이를 보였을 뿐이다.

이상을 요약하면, 단독발화보다 동시발화에서 발화속도가 확연하게 느려졌다. 그리고 동시발화의 유무가 발화속도에 미치는 영향은 남녀 두 화자집단 간의 차이가 발화속도에 미치는 영향보다 크게 나타났다.

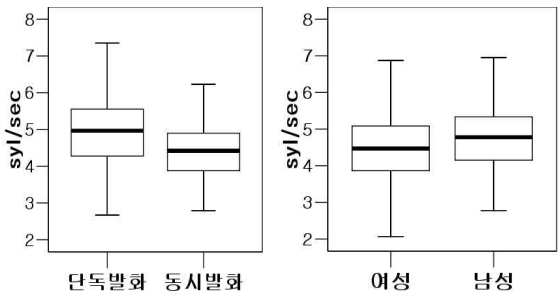

그림 5. 동시발화 유무(단독발화 vs. 동시발화)와 남녀 화자에 따른 발화속도 상자도표

Figure 5. Boxplot of average speech rate (syl/sec) of different speaking modes (solo vs. synchronized) and gender (male vs. female).

3.2 발화문장 내 피치포인트 간 억양자질 분석

동시발화 유무와 남녀 두 화자집단을 모수요인으로 하고 추출된 억양궤적으로부터 얻은 피치포인트 자질들(Pt, Pr, Ps, Pd)을 종속변수로 하는 MANOVA(다변량 분산분석)를 시행했다. 동시발화의 유무는 Pt[F(1, 380.507), p<0.001, η_P^2=0.017], Pr[F(1, 483.186), p<0.001, η_P^2=0.022], Ps[F(1, 76.313), p<0.001, η_P^2=0.004], Pd[F(1, 687.460), p<0.001, η_P^2=0.031] 모두에서 차이를 나타냈다.

<그림 6>을 보면, 단독발화에서 동시발화로 바뀜에 따라 Pt(피치포인트 이동시간), Pr(피치포인트 피치변위), Ps(피치포인트 기울기), Pd(피치포인트 이동거리) 등이 모두 감소하는 것을 알 수 있다. 이것은 동시발화에서 발화속도가 감소하고 피치포인트 개수가 증가해서 피치포인트들의 활동이 둔화되고 피치포인트들 간의 간격이 줄어들었기 때문으로 보인다(<그림 4> 참조).

이것은 빠른발화에서 발화속도 증가로 인해 피치포인트 개수가 줄어들었다는 연구(이서배 2014b)와 함께 발화속도와 피치포인트 개수 간에 음의 상관관계(반비례)를 생각할 수 있게 하는 결과이다.

동시발화 유무와 남녀 화자간의 상호작용은 Pt[F(1, 0.063), p=0.802], Pr[F(1, 2.944), p=0.086], Ps[F(1, 0.218), p=0.641], Pd[F(1, 1.625), p=0.202] 중 어떤 억양자질에서도 나타나지 않았다. 그래서 남녀를 구분해서 동시발화 유무를 따로 분석하지는 않았다.

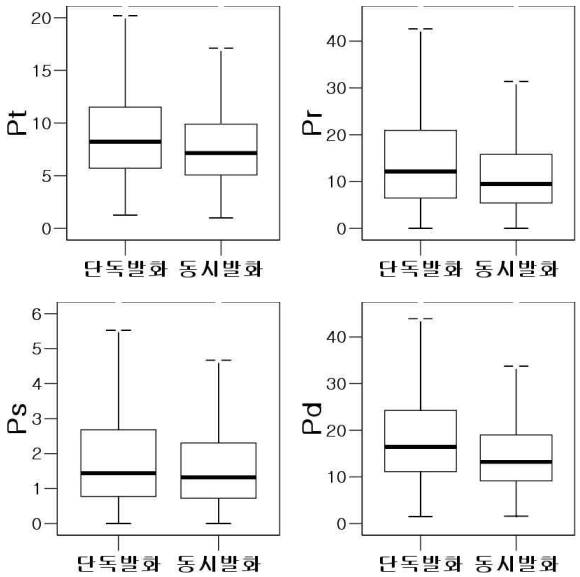

그림 6. 동시발화 유무(단독발화 vs. 동시발화)에 따른 피치포인트 간 억양자질 값의 상자도표

Figure 6. Boxplot of pitch point values according to different speaking modes (solo vs. synchronized).

3.3 문장수준 전체 억양자질 분석

동시발화 유무와 남녀 두 화자집단을 모수 요인으로 하고 문장수준의 억양자질들(Sr, Ss, Sp, Sd, Max, Min, MaxNr, MinNr)을 종속 변수로 하는 MANOVA를 시행했다.

3.3.1 동시발화 유무에 따른 분석

동시발화 유무는 Sr[F(1, 76.082), p<0.001, η_P^2= 0.040], Ss[F(1, 11.342), p<0.01, η_P^2=0.006], Sd[F(1, 83.689), p<0.001, η_P^2=0.043], MaxNr[F(1, 77.822), p<0.001, η_P^2=0.040], MinNr[F(1, 14.566), p<0.001, η_P^2=0.008], Min[F(1, 39.990), p<0.001, η_P^2=0.021]에서 유의미한 것으로 나타났다. 그러나 Sp[F(1, 3.428), p=0.064], Max[F(1, 2.257), p=0.133]에서는 유의미한 차이가 없었다.

<그림 7>을 보면 동시발화에서 MaxNr, MinNr, Sr, Ss는 줄어들고 Min과 Sd는 증가하는 것을 볼 수 있다. MaxNr, MinNr은 Max와 Min이 한 문장의 중앙값(Mid)으로부터 떨어진 거리를 의미하는데 중앙값과의 거리가 클수록 큰 값을 가지게 된

다. 동시발화에서 MaxNr, MinNr이 줄어들었다는 것은 그만큼 중앙값과의 거리가 좁아졌다는 것을 의미하고 결국 문장 피치 변위인 Sr도 줄어들었다는 이야기이다. Ss의 경우 'Ss = Sp / Sd'로 구해지는데 Sp는 의미 있는 차이가 없으므로 결국 Sd의 차이로 인해 Ss가 결정된다고 볼 수 있다. 즉, 동시발화에서 발화시간이 늘어남으로 Sd가 증가했고 이에 따라 Sd와 반비례 관계를 가지는 Ss는 감소하게 된 것이다.

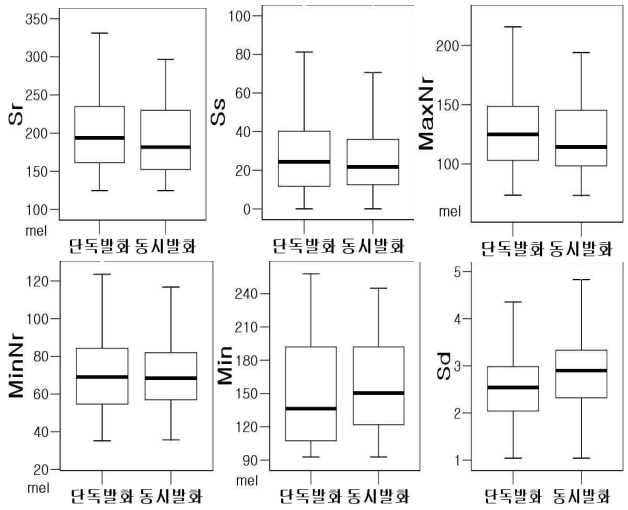

그림 7. 동시발화 유무(단독발화 vs. 동시발화)에 따른 문장수준 억양자질 값의 상자도표

Figure 7. Boxplot of sentence level intonation feature values according to different speaking modes (solo vs. synchronized)

동시발화에서도 변화를 보이지 않은 Max와는 달리 Min값은 동시발화에서 증가하는 것으로 나타났다. 그러므로 문장 피치 변위의 감소는 Max의 변화보다는 Min값의 상승이 주된 작용을 한 것으로 보인다. 이것은 빠른발화의 연구에서 Min은 거의 변화가 없었던 반면 Max의 값이 작아지므로 문장피치변위의 폭이 줄어들었다는 보고(이서배, 2014b)와 비교할 때 흥미로운 결과다. 즉, 빠른발화와 마찬가지로 본 연구의 동시발화에서도 문장전체의 피치범위가 줄어들었는데 빠른발화에서는 Max의 하락이, 동시발화에서는 Min의 상승이 문장피치범위의 감소를 주도한 것이다. 동시발화에서 나타나는 문장피치변위 감소현상이 전반적인 발화속도 감소에 의한 것인지 동시발화 화자들 간의 리듬동조화(entrainment)현상에 의한 것인지는 향후 추가적인 연구가 필요할 것이지만 이러한 결과는 동시발화에서 문장전체의 피치범위가 상당히 줄어들었다는 기존연구(Cummins, 2000; Cummins & Roy, 2001)와 맥을 같이하고 있다.

또 다른 흥미로운 결과는 단독발화에서 Mid, MaxNr, MinNr의 값이 큰 값일수록 단독발화의 Mid, MaxNr, MinNr에서 동시발화의 Mid, MaxNr, MinNr을 뺀 값(단독발화-동시발화)이 커지고 단독발화의 Mid, MaxNr, MinNr의 값이 작은 값일수록 단독발화의 Mid, MaxNr, MinNr에서 동시발화의 Mid, MaxNr, MinNr을 뺀 값(단독발화-동시발화)이 작아진다는 것이다. <그림 8>의 수직축에서 '단독발화-동시발화'의 값이 0이 될 때('단독발화-동시발화'의 값이 0이 되는 지점을 수평 실선으로 표시했다), 단독발화의 억양자질 값과 동시발화의 억양자질 값이 같아진다. 그리고 '단독발화-동시발화'의 값이 0보다 클수록 단독발화의 억양자질 값은 크고 동시발화의 억양자질 값은 작다는 것을 의미한다. 그리고 '단독발화-동시발화'의 값이 0보다 작을수록 단독발화의 억양자질 값은 작고 동시발화의 억양자질 값은 크다는 것을 의미한다.

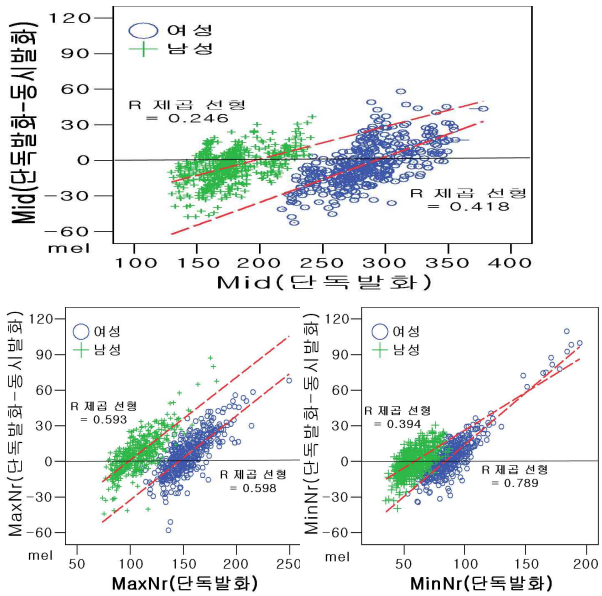

그림 8. 동시발화 유무(단독발화 vs. 동시발화)에 따른 남녀의 Mid, MaxNr, MinNr 변화

Figure 8. Change of Mid, MaxNr, MinNr of men and women according to different speaking modes (solo vs. synchronized)

그러므로 <그림 8>의 수직축에서 '단독발화-동시발화'의 값이 0보다 큰 경우, 단독발화에서 값이 큰 Mid, MaxNr, MinNr 일수록(수평축) '단독발화-동시발화'(수직축)의 값이 커진다. 즉, 동시발화에서 Mid, MaxNr, MinNr의 값이 더 작아지는 것이다. 이러한 현상은 보통속도의 발화에서 큰 Max일수록 빠른 속도의 발화에서 Max가 더 많이 감소하는(보통속도 발화의 Max를 기준으로) 경향이 있는 것을 보여준 기존연구(이서배, 2014b, Fougeron & Jun, 1998)와 비슷한 현상이다. 그리고 '단독발화-동시발화'의 값이 0보다 작은 경우, 단독발화에서 Mid, MaxNr, MinNr의 값이 작을수록 동시발화에서 Mid, MaxNr, MinNr의 값이 더 커지는 경향을 볼 수 있다.

이와 같이 단독발화와 동시발화 간에 나타나는 Mid, MaxNr, MinNr 변화의 반비례적 선형관계를 기존연구(Fougeron & Jun, 1998)에서 말하는 포화효과(Saturation Effect)의 관점에서 살펴보면, 단독발화의 Mid, MaxNr, MinNr 값이 동시발화에서 도달해야하는 목표 Mid, MaxNr, MinNr이 일정 수준으로 정해져 있

어서 단독발화에서 큰(작은) Mid, MaxNr, MinNr은 동시발화에서 많이 감소(증가)하지만 이미 목표 값에 근접한 작은(큰) Mid, MaxNr, MinNr(단독발화의)은 동시발화에서 작게 감소(증가)할 수밖에 없다는 해석이 가능해 진다. 이러한 경향은 남성보다는 여성에게서 더 강하게 나타나고 있는데(<표 1>, <표 2>, <표 3>참조) 이것은 남성보다 문장수준의 피치변위 폭이 큰 여성이 동시발화에서 더 많은 변화의 여지를 가지기 때문으로 추정된다(<그림 9>의 Sr값 참조).

표 1. 남녀에 따른 '단독발화 Mid'와 '단독발화 Mid - 동시발화 Mid'간의 상관계수

Table 1. Correlation coefficients between 'Mid of solo speech' and 'Mid of solo speech minus Mid of synchronized speech' for men and women

	여성	남성
Pearson 상관계수	0.646(**)	0.496(**)
유의확률 (양쪽)	p<0.001	p<0.001
발화문장 수	396	528

표 2. 남녀에 따른 '단독발화 MaxNr'과 '단독발화 MaxNr - 동시발화 MaxNr'간의 상관계수

Table 2. Correlation coefficients between 'MaxNr of solo speech' and 'MaxNr of solo speech minus MaxNr of synchronized speech' for men and women

	여성	남성
Pearson 상관계수	0.773(**)	0.770(**)
유의확률 (양쪽)	p<0.001	p<0.001
발화문장 수	396	528

표 3. 남녀에 따른 '단독발화 MinNr'과 '단독발화 MinNr - 동시발화 MinNr'간의 상관계수

Table 3. Correlation coefficients between 'MinNr of solo speech' and 'MinNr of solo speech minus MinNr of synchronized speech' for men and women

	여성	남성
Pearson 상관계수	0.888(**)	0.628(**)
유의확률 (양쪽)	p<0.001	p<0.001
발화문장 수	396	528

3.3.2 남녀 두 화자집단에 따른 분석

동시발화의 유무와 남녀에 따른 상호작용은 Sr[F(1, 5.996), p<0.05, η_P^2= 0.003]과 MinNr[F(1, 19.238), p<0.001, η_P^2=0.010]에서만 나타났으므로 Sr과 MinNr은 남녀 각각의 경우를 나누어 분석할 필요가 있다. 그래서 남녀 각각에 따라 동시발화의 유무를 모수요인으로 하고 Sr과 MinNr을 종속 변수로 하는 MANOVA를 시행했다.

여성의 경우, 동시발화의 유무가 Sr[F(1, 48.450), p<0.001, η_P^2= 0.058], MinNr[F(1, 19.431), p<0.001, η_P^2=0.024] 모두에 유의미한 차이를 주는 것으로 나타났다. 그러나 남성의 경우, 동시발화의 유무가 Sr[F(1, 25.374), p<0.001, η_P^2= 0.024]에서는 유의미한 차이를 주었지만 MinNr[F(1, 0.308), p=0.579]에서는 유의미한 차이를 주지 않았다. <그림 9>에서와 같이 남녀 모두 단독발화에 비해 동시발화에서 Sr이 감소하는 반면 MinNr은 여성에게서만 이러한 감소현상이 나타나는 것을 볼 수 있다.

정리하면, 동시발화의 Sr은 남녀 두 화자집단에서 의미 있는 감소를 보였지만 효과크기로 보면 남성의 Sr(η_P^2= 0.024)보다 여성의 Sr(η_P^2= 0.058)이 더 큰 변화를 보였다. 그리고 MinNr의 감소는 여성에게는 나타났지만 남성에게는 나타나지 않았다. 이것은 남성보다 더 큰 피치범위를 가지는 여성의 억양이 남성의 억양보다 변화의 여지가 더 크고 민감했기 때문으로 추정된다.

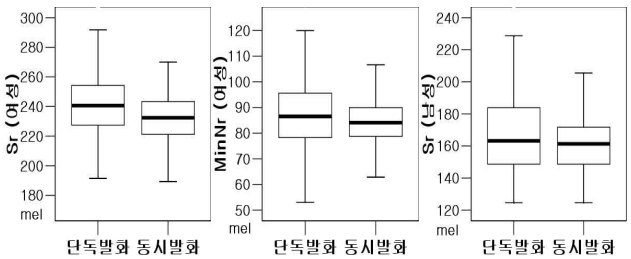

그림 9. 동시발화 유무(단독발화 vs. 동시발화)에 따른 남녀의 문장수준 억양자질 값의 상자도표

Figure 9. Boxplot of sentence level intonation feature values of men and women according to different speaking modes (solo vs. synchronized).

3.3.3 화자 변별력을 가지는 억양자질

동시발화에서는 두 화자가 서로 발화 타이밍을 맞추기 때문에 억양곡선도 비슷해지는 경향이 생긴다(Cummins, 2000). 이로 인해 두 화자의 특성을 구분하는 일(화자인식)이 더 힘들어진다. 그럼에도 불구하고 어떤 억양자질(억양자질 X)이 A화자의 단독발화와 동시발화를 구별해 주는 것보다 A화자의 동시발화와 B화자의 동시발화를 더 잘 구별해 준다면 그 억양자질은 화자들 간의 변별력이 높을수록 좋은 화자인식시스템에서 효과적으로 사용될 수 있을 것이다.

$$|X_{Synch\,A} - X_{Solo\,A}| < |X_{Synch\,A} - X_{Synch\,B}| \quad (1)$$

(Cummins et al., 2006)

수식 (1)에서, $X_{Synch\,A}$는 A화자의 동시발화에서 추정한 자질 X, $X_{Synch\,B}$는 B화자의 동시발화에서 추정한 자질 X, $X_{Solo\,A}$는 A화자의 단독발화에서 추정한 자질 X를 의미한다.

그러므로 $|X_{Synch\,A} - X_{Solo\,A}|$는 한 화자 내 변이(intra-speaker variation)에 해당하고 $|X_{Synch\,A} - X_{Synch\,B}|$는 화자들 간의 변이(inter-speaker variation)에 해당하는 셈이다. 수식 (1)을 본 연구의 억양자질들에 적용한 결과, Pt, Pr, Ps, Pd, Sr, Ss, Sp, Sd, Max, Min, MaxNr, MinNr, Mid 중에서 Max, Min, Mid가 화자인식에 유용한 변별력을 가지는 것으로 나타났다.

4. 결 론

영어의 단독발화와 동시발화의 억양분석을 위해 자동 억양궤적 추출 알고리듬인 Momel을 이용해 억양자질들의 값을 구했다. 동시발화는 단독발화보다 발화속도가 현저히 저하되었고 이것은 남녀 간의 발화속도 차이보다도 큰 변화였다.

발화문장 내 피치포인트 간 억양자질 분석에서 동시발화가 단독발화보다 Pt(피치포인트 이동시간), Pr(피치포인트 피치변위), Pd(피치포인트 이동거리), Ps(피치포인트 기울기)의 값이 더 작아 동시발화에서 피치포인트들의 역동성이 떨어지는 현상이 나타났는데 이것은 동시발화에서 발화속도가 늦어졌고 피치포인트 개수가 증가해서 피치포인트들의 활동이 둔화되고 그 영역이 좁아진 때문으로 해석된다. 한편, 동시발화 유무와 남녀 두 화자집단의 상호작용은 나타나지 않았다.

발화문장 전체 억양자질 분석을 통해 동시발화에서 MaxNr(Max와 중앙값과의 차이), MinNr(Min과 중앙값과의 차이), Sr(문장 피치변위), Ss(문장 피치기울기)는 감소했고 Min(피치 최솟값)과 Sd(첫 피치포인트에서 마지막 피치포인트까지의 시간)는 증가하는 것을 알 수 있었다. MaxNr, MinNr의 감소는 Sr의 감소를 의미하고 Sr의 감소는 Max의 변화 보다는 Min의 변화(증가)에 의한 것으로 나타났다. Ss는 Sd와 반비례하므로 Sd의 증가에 따라 감소하였다.

단독발화와 동시발화 간에 Mid, MaxNr, MinNr 변화의 반비례적 선형관계도 나타났는데 포화효과(Saturation Effect)에 의하면 단독발화의 Mid, MaxNr, MinNr 값이 동시발화에서 도달해야하는 목표 Mid, MaxNr, MinNr의 값이 일정 수준으로 정해져 있어서 단독발화에서 큰(작은) Mid, MaxNr, MinNr은 동시발화에서 많이 감소(증가)하지만 이미 목표 값에 근접한 작은(큰) Mid, MaxNr, MinNr(단독발화의)은 동시발화에서 작게 감소(증가)할 수밖에 없다는 설명이 가능하다. 이러한 경향은 남성보다는 여성에게서 더 높은 상관관계로 나타났다. 이것은 여성이 남성보다 피치변위의 폭이 더 크므로 동시발화에 민감하게 반응할 여지가 더 크기 때문으로 추정된다.

본 연구의 결과에 나타난 바와 같이 동시발화에서 억양의 움직임이 완만해지는 현상은 화자 내 변이의 감소를 의미하고 이것은 코퍼스기반 음성합성(Concatenative Speech Synthesis)에 유용하게 적용 될 수 있을 것이다. 즉, 이어 붙는 두 유닛(unit)의 음향특질이 비슷할수록 음성신호처리로 인한 왜곡이 최소화될 수 있으므로 유닛간의 시간적, 억양적 변화를 줄여 음향특질 차이를 감소시키는 동시발화기법은 코퍼스기반 음성합성의 음성코퍼스 구축방법으로 고려해 봄직하다(Cummins & Roy, 2001). 또한 화자 내 변이보다 화자들 간의 변이를 더 크게 반영하는 억양자질의 화자변별력이 크다고 볼 때, Max, Min, Mid가 화자인식에 유용한 억양자질로 보인다. 그리고 남성보다 여성의 억양변화가 더 크고 민감한 점으로 보아 화자인식, 화자검증, 감정인식 등과 같이 F0를 입력자질들 중의 하나로 사용하는 시스템에서는 남녀 가중치 차별화를 시도해 볼 만할 것이다.

감사의 글

I would like to express my special appreciation and thanks to Dr. Beach and Mr. Beach. I would never have been able to finish this research paper without their tremendous support. I would like to thank them for allowing me to use their language lab and encouraging my research. Their advice on both my research as well as my career has been priceless.

참고문헌

Banzina, E., Hewitt, L. & Dilley, L. (2014). Using synchronous speech to facilitate the acquisition of English rhythm: A small scale study. E-JournALL, EuroAmerican *Journal of Applied Linguistics and Languages,* 1(1), 69-84.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*. 5:9/10, 341-345.

Campione, E. (2001). *Etiquetage prosodique semi-automatique de corpus oraux : algorithmes et méthodologie.* Thèse de doctorat. Aix-en-Provence: Université de Provence.

Cummins, F. (2000). Prosodic characteristics of synchronous speech in Puppel, S. and Demenko, G., editors, *Prosody 2000: Speech Recognition and Synthesis*, Adam Mickiewicz University, Krakow, Poland, 45-49.

Cummins, F. & Roy, D. (2001). Using synchronous speech to minimize variability in pause placement. *Proceedings of the Institute of Acoustics,* Stratford-upon-Avon, 23 (3), 201-206.

Cummins, F. (2003). Practice and performance in speech produced synchronously. *Journal of Phonetics,* 31(2), 139-148.

Cummins, F., Grimaldi, M., Leonard, T. & Simko, J. (2006). The CHAINS corpus: Characterizing individual speakers. *Proceedings of SPECOM 2006,* 431–435, St. Petersburg, RU.

Cummins, F., Li, C. & Wang, B. (2013).Coupling among speakers

during synchronous speaking in English and Mandarin. *Journal of Phonetics.* 41(6), 432-441.

Forsell, M. (2007). *Acoustic correlates of perceived emotions in speech.* MS Thesis, KTH, Royal Institute of Technology, Stockholm, Sweden.

Fougeron, C. & Jun, S. (1998). Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics*, 26, 45-69.

Hirst, D., Cristo, A. & Espesser, R. (2000). Levels of representation and levels of analysis for intonation. in M. Horne (ed) Prosody : *Theory and Experiment.* Kluwer Academic Publishers, Dordrecht. 51-87.

Hirst, D. (2007). A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation. *Proceedings of the XVIth International Conference of Phonetic Sciences*, Saarbrucken, 1233-1236.

Kim, M. & Nam, H. (2012). Speech Rate Variation in Synchronous Speech. *Phonetics and Speech Sciences,* 4(4), 19-27.
(김미란, 남호성 (2012). 동시발화에 나타나는 발화 속도 변이 분석. 말소리와 음성과학, 4(4), 19-27.)

Looze, C. & Hirst, D. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation, *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 135-138.

Oh, J. H. (2014a). A Study of methods of standardization for Korean intonation curve. *Korean Linguistics,* 62, 395-420.
(오재혁 (2014a). 한국어 억양 곡선의 정규화 방안에 대한 연구. 한국어학, 62, 395-420.)

Oh, J. H. (2014b). A study of intonation curve slopes in Korean spontaneous speech. *Phonetics and Speech Sciences,* 6(1), 21-30.
(오재혁, (2014b). 자유 발화 자료에서 나타나는 한국어 억양 곡선의 기울기 특성에 대한 연구. 말소리와 음성과학, 6(1), 21- 30.)

Quene, H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics.* 35, 353-362.

Yi, S., & Kim, S. (2011). A study on low pitch accent produced in different locations in English sentences. *Phonetics and Speech Sciences*, 3(4), 63-70.
(이서배, 김수정 (2011). 영어 문장 내 상이한 위치에 나타난 저성조 피치 액센트 연구, 말소리와 음성과학, 3(4), 63-70.)

Yi, S. (2014a). An acoustical analysis of emotional speech using close-copy stylization of intonation curve, *Phonetics and Speech Sciences,* 6(3), 131-138.
(이서배, (2014a). 억양의 근접복사 유형화를 이용한 감정음성의 음향분석, 말소리와 음성과학, 6(3), 131-138.)

Yi, S. (2014b). An acoustical analysis of speech of different speaking rates and genders using intonation curve stylization of English, *Phonetics and Speech Sciences,* 6(4), 79-90.
(이서배, (2014b). 영어의 억양 유형화를 이용한 발화 속도와 녀 화자에 따른 음향 분석, 말소리와 음성과학, 6(4), 79-90.)

• **이서배 (Yi, So Pae)**
창원대학교 영어영문학과
(641-773) 경남 창원시 의창구 사림동
Tel.: 055-540-5466 Fax: 055-540-5465
Cell: 010-5555-6305
Email: sopaeyi@pusan.ac.kr

부록

Daniel Hirst의 praat 스크립트들(2008, July)을 통합하고 수정해서 대용량의 음성파일들을 한 번에 처리하고 본 연구에 필요한 값들을 얻을 수 있도록 했다. 억양궤적을 추출하는 momel 알고리듬 실행화일 'momel_win.exe' (Momel저자 홈페이지에서 구할 수 있음)가 본 스크립트와 같은 폴더에 있어야 한다.

```
########    Beginning of script    ###########
sound_extension$ = ".wav"
pitch_extension$ = ".hz"
momel_extension$ = ".momel"
pitchTier_extension$ = ".pitchtier"
momel1$ = "momel_win.exe"
momel_parameters$ = "30 60 750 1.04 20 5 0.05"
minimum_f0 = 60
maximum_f0 = 750
pitch_step = 0.01
Create Strings as directory list... dirListObj 'inputFolder$'/*
numFolders = Get number of strings
#모든 화자들의 개별 폴더를 하나씩 처리함
for iFolder from 1 to numStrings
   select Strings dirListObj
   dirName$ = Get string... iFolder
   Create Strings as file list... fileListObj 'inputFolder$'/'dirName$'/
*'sound_extension$'
   Sort
   numFiles = Get number of strings
     for iFile to numFiles
        select Strings fileListObj
        file$ = Get string... iFile
        Read from file... 'inputFolder$'/'dirName$'/'file$'
        prefix$ = file$-sound_extension$
        select Sound 'prefix$'
```

```
        duration = Get total duration
        sound_file$ = prefix$+sound_extension$
        pitch_file$ = prefix$+pitch_extension$
        momel_file$ = prefix$+momel_extension$
        momelPitchTier$ = prefix$+pitchTier_extension$

# 최적의 f0 최대값, f0 최소값은 디폴트값으로 구한 피치에서 1st & 3rd quartiles로 구할 수 있고 그 공식은 다음과 같다(Looze & Hirst, 2008).
# f0 max = 1.5 * q3 (3번째 quartile); f0 min = 0.75 * q1 (1번째 quartile)
# rounded to higher (resp. lower) 10
            select Sound 'prefix$'
            To Pitch... 'pitch_step' 'minimum_f0' 'maximum_f0'
            .q75 = Get quantile... 0.0 0.0 0.75 Hertz
            .q25 = Get quantile... 0.0 0.0 0.25 Hertz
            max_f0 = 10*ceiling((1.5*.q75)/10)
            min_f0 = 10*floor((0.75*.q25)/10)
            select Sound 'prefix$'
            myPitch = To Pitch... 'pitch_step' 'min_f0' 'max_f0'
            nValues = Get number of frames
            myMatrix = To Matrix
            Transpose
            Write to headerless spreadsheet file... 'pitch_file$'
#'momel_win.exe'을 이용해서 억양곡선 추정
    system 'momel1$' > 'momel_file$' 'momel_parameters$' < 'pitch_file$'
            myStrings = Read Strings from raw text file... 'momel_file$'
            nStrings = Get number of strings
        for iString from 1 to nStrings
            select myStrings
            string$ = Get string... iString
            ms = extractNumber(string$,"")
            secs = ms/1000
            f0 = extractNumber(string$, " ")
            if ms = undefined
             printline String ['iString'] ('string$') doesn't contain a number
            else
                if f0 > max_f0
                   f0 = max_f0
                elsif f0 < min_f0
                    f0 = min_f0
                endif
                time = secs
                if time <0
                    time = 0
                elsif time > duration
                    time = duration
                endif
                pointT[iString] = time
                pointP[iString] = f0
                tmpF0[iString] = f0
            endif
         endfor #iString

# 추정된 f0들에서 bubble sort를 이용해 중앙값 찾기
            i = nStrings
            j = 1
            while (i > 0)
               while (j < i)
                  if tmpF0[j] > tmpF0[j+1]
                     tmp = tmpF0[j]
                     tmpF0[j] = tmpF0[j+1]
                     tmpF0[j+1] = tmp
                  endif
                  j = j + 1
               endwhile
               i = i - 1
               j = 1
            endwhile
            midPoint = round(nStrings/2)
            midF0 = tmpF0[midPoint] # f0 중앙값
            for iString from 1 to nStrings-1
               nowT = pointT[iString] #현 피치포인트 시각
               nextT = pointT[iString+1] #다음 피치포인트 시각
               firstT = pointT[1] #최초 피치포인트 시각
               lastT = pointT[nStrings] #마지막 피치포인트 시각
               nowP = pointP[iString] #현 피치포인트 피치
                nextP = pointP[iString+1] #다음 피치포인트 피치
               firstP = pointP[1] #최초 피치포인트 피치
               lastP = pointP[nStrings] #마지막 피치포인트 피치
            endfor      #변수 출력 루틴 생략...
      endfor    #메모리확보를 위해 로딩된 파일 제거 루틴 생략
endfor #iFolder
########    End of script    ############
```