제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

대화체 음성언어번역 시스템 개발

박준, 이영직, 양재우

대전광역시 유성구 가정동 161, 한국전자통신연구원 휴먼인터페이스연구부

Spontaneous Speech Translation System Development

Jun Park, Youngjik Lee, and Jaewoo Yang

Human Interface Department, ETRI, 161 Kajong-dong, Yusung, Taejon

junpark@etri.re.kr

요약: 본 논문에서는 ETRI 에서 개발 중인 대화체 음성언어번역 시스템에 대하여 기술한다. 현재, ETRI 는 음성언어번역 국제 공동 연구콘시시움인 C-STAR에 핵심참가기관으로 참여하여, 한일, 한영 음성언어번역 시스템을 개발하고 있으며 1999년 국제 공동시험을 계획하고 있다. 최근의 연구 진행상황을 간추려면, 먼저 음성인식분야에서 유무성음 및 묵음정보를 미리 추출하여 이를 탐색에 활용하였으며, 음향모델 규모의 설정을 위한 교차 엔트로피 기반 변이음 군집화 알고리즘이 구현되었다. 또한 대상어휘의 확장을 위하여 의사형태소의 개념을 도입하였다. 언어번역분야에서는 이전과 같이 개념기반의 번역을 시도하고 있으며, C-STAR 회원기관과 공동으로 중간언어 규격을 정의하고 있다. 음성합성 분야에서는 훈련형 합성기를 개발하여 합성데이타베이스 구축기간을 현저하게 줄였다.

1. 서론

음성언어번역 시스템이란 서로 다른 언어를 사용하는 사람 간에 대화가 가능하게 하는 궁극적인 통신 도구이다. 이러한 시스템을 개발하기 위하여는 먼저 사람이 발성한 소리를 문자로 나타내는 음성인식기술과, 이를 같은 의미를 갖도록 다른 언어로 표현하는 언어번역기술, 그리고 문자로 표현된 문장을 읽어 주는 음성합성기술 등의 요소기술이 확보되어야 한다. 그간 상상 속에서만 존재하던 음성언어번역 시스템은 최근 컴퓨터 및 반도체 기술의 급속한 발전을 기반으로 각 요소기술의 완성도가 확보됨에 따라 머지 않은 장래에 상용화가 될 것으로 전망되고 있다.

특히, 1993년 일본 ATR-ITL 연구소, 미국의 카네기멜런대학교 (CMU), 독일의 지멘스가 공동으로 1,500 어휘급 학회등록 작업영역에서 동작하는 영어, 독일어, 일본어간 낭독체 음성언어번역 시스템을 공동 시연함으로써 음성언어번역 시스템의 실현 가능성을 입증한 바 있다. 이러한 시연의 성공에 힘입어, 처리 대상 대화를 정확하게 발성하는 낭독체로부터 자연스럽게 발성하는 대화체로 확장하고, 작업영역도 여행계획이라는 보다 포괄적인 영역으로 확대하며, 다양한 언어를 수용하는 국제 음성언어번역 연구 콘소시엄인 C-STAR (Consortium for Speech Translation Advanced Research) 가 1995년 발족하였다. 이 콘소시엄에는 ATR, CMU, 지멘스 외에도 이탈리아의 IRST, 프랑스의 CLIPS 그룹, LIMSI, 미국의 MIT, MIT Lincoln Lab., AT&T, 영국의 SRI 등 음성언어번역 분야의 첨단 연구기관이 대거 참여하고 있으며, ETRI 도 핵심그룹의 일원으로 참여하고 있다. 음성언어번역시스템은 적어도 두개 이상의 언어를 처리하여야 하므로 그 성격상 국제간 공동연구 개발이 효율적이다. 따라서 그림 1에 보는 바와 같이 ETRI 는 전체 시스템의 한/영, 한/일 대화체 음성언어번역 시스템을 개발하고 있으며, 일본의 ATR-ITL, 미국의 CMU 가 개발하고 있는 일한, 영한 시스템과 연동하여, 1999년에 한일, 한영 음성언어번역 공동 시험을 실시할 예정이다.

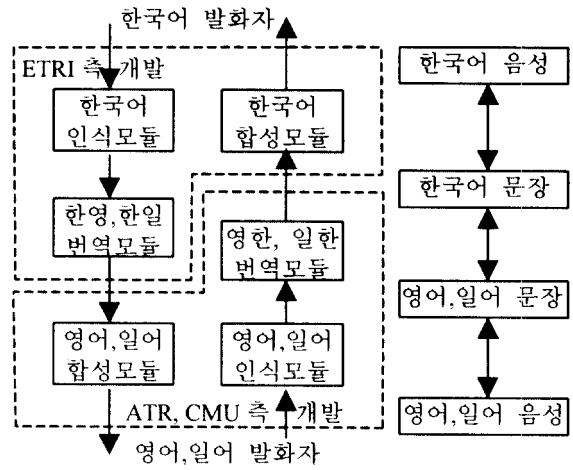

그림 1. 음성언어번역 시스템 구성도

본 논문에서는 현재 ETRI 에서 개발 중인 음성언어번역 시스템의 현황을 기술적인 내용을 중심으로 소개한다. 먼저 시스템 구축을 위하여 수집한 음성데이타에 대하여 설명하고, 이어서 각 모듈 별로 특징 및 최근 개발 상황, 그리고 앞으로의 과제를 설명한다.

2. 음성데이타

음성 관련 연구의 성공의 첫걸음은 데이타의 확보 여부에 달려 있다고 할 만큼 음성데이타는 중요한 역할을 차지한다. ETRI 에서도 시스템 개발을 위하여 다양한 음성데이타를 수집하였다. 먼저, 음향모델의 효율적인 훈련을 위하여 표 1 과 같이 트라이폰 단위로 음소의 분포를 고려하는 POW (Phonetically Optimized Word) 단어 세트를 추출하여 데이터를 수집하였으며, 일부 데이터에 대하여 음소별로 분절표기를 하였다.

표 1. POW 음성데이타 사양

항목	내용
POW 어휘 목록	3848 개 한글 완성형 파일
단어음성 녹음 DB	총 38,480 단어 1Gbytes
단어경계 정보 파일	38,480 단어 남성음: 5 set, 여성음: 5 set (각 set 은 3,848 개 단어로 구성되며, 8 명이 나누어 발성함)
발음사전	3,848 개 어휘에 대한 음소 발음 기호열
음소경계 정보 파일	총 19,240 단어 남성음: 3 set, 여성음: 2 set

대화체 음성 데이터는 시간약속 작업 영역과 이행계획 작업 영역을 설정하고 대화음성을 수집하였다. 대화체 문장에서는 그림 2 의 예에서 보는 바와 같이 반복, 축약, 생략 등 비정형문이 빈번히 발생한다. 이러한 특성을 지닌 이러한 데이터를 수집함에 있어 모의 상황을 설정하고 발화자로 하여금 자신의 역할을 잘 숙지시켜 자연스러운 대화 음성을 얻어야 한다. 그런데, 화자독립 음향모델을 훈련시키고, 언어모델 구축에 필요한 다양한 발화문을 얻기 위하여 보다 많은 사람의 발화자로 하여금 각각 소량의 분량을 발성하게 하여야 하므로 모의 대화자의 훈련에 어려움이 따랐다. 즉, 모의대화자에 주어지는 정보가 적으면 대화진행이 원활하지 못하였고, 사전 정보가 너무 많이 주어져도 상대방이 알려주어야 하는 정보를 미리 말하거나, 대화의 진행이 지나치게 빠르게 되고 발화문도 정형문에 가까워져 대화체의 특징이 사라지는 현상이 발생하였다. 이러한 문제에 대하여 각 발화자에 사전 상황 설명을 하고, 실제 대화의 순서와 각자의 역할에서 필요한 정보가 명시되어 있는 도표를 사전에 제공하고 상대방으로부터 얻어야 하는 정보는 빈칸으로 표기하여 대화를 통하여 적도록 유도하므로써 자연스러운 대화 데이타를 얻었다. 그리고, 같이 대화하는 상대방이 평소에 잘 알고 있는 사람이거나, 대화 상황이 마주보며 말할 경우에는 수집되는 발화문장의 구조가 더욱 심하게 왜곡되는 반면, 컴퓨터 화면을 통하여 대화할 때나, 중간에 동시통역사를 두고 외국인과 대화할 때는 발화자들이 정형문에 가까운 문장을 발화하는 현상이 나타났다. 이는 음성언어번역 시스템이 실제 사용되는 상황을 고려하면 시사하는 바가 크다고 할 수 있다.

수집된 데이터는 ETRI 에서 정하는 전사규칙에 준하여 전사되었으며, 이 규칙에는 잡음, 간투사, 반복발성, 사투리 등에 대한 규정을 포함하고 있다. 본 시스템 개발을 위하여 수집된 데이터는 대학 및 업체에 배포하고 있는데, POW 데이터는 '98 년 2 월에 공개하였으며, 검증이 완료된 시간약속 작업 영역 대화체 데이터도 1998 년말 배포할 계획이다.

갑: 에 두 둘째주 십사일날 오후쯤에 시간이 나면은 테니스 쳐가지고 어 가족대항 테니스를 쳐가지고 어 저녁때 식사나 한껏 하려고 그러는데요.

을: 아 십사일날 저도 뭐 뭐 특별한 일은 아니구요. 어 간단히 연구실쪽에서 어 뭐야 체육대회 한다고 뭐 테니스가 그렇게 뭐 쏟 하고 그런다고 그래서 선약이 있어가지고 십사일은 좀 안되겠는데요.

그림 2. 시간약속 영역 수집데이타 예

3. 음성인식모듈

ETRI 의 음성인식기는 HMM 기반 인식기[1]로서 처음 45 대화(대화당 평균 10 발화)의 음성데이타를 사용하여 문맥독립 47 개 음소모델(6 개 잡음 포함), 각 음소 당 50 개 코드워드, 16 차원 Melscale FFT 로 특징을 추출하는 기반시스템으로부터 현재까지의 구현 과정은 표 2 와 같다. 이어서 최근의 연구결과인 원격음성입력기, 지식기반 탐색, 교차 엔트로피 기반 변이음 군집화, 대용량 인식을 위한 의사형태소 등을 아래에 기술한다.

현재 발표되는 대부분의 음성인식시스템은 하나의 마이크를 이용하여 음성을 입력하는 방법을 채택하고 있다. 그러나, 하나의 마이크로 음성을 입력받는 경우에는 음성입력 시에 마이크의

표 2. 인식기 구현단계와 성능지표

추가내용	인식률(%)	감소율(%)	비고
기반시스템	40.7	-	시간약속 45 대화
135 대화로 증강	33.7	-17.3	45 대화에서, 사전크기 증강
16 차원 LDA	58.1	36.8	16 차원 dMel 추가하여 LDA 적용
훈련반복	69.5	27.2	이상 훈련데이타 대상
평가데이타 분리	29.4	-	이하 평가데이타 대상
전사수정	39.3	14.0	등가발성 반영
문맥종속음향모델	46.0	11.0	1,940 triphone
데이터증강	47.3	2.4	304 대화로
변이음 분류	54.0	12.7	1,295 개 변이음
PLP 특징도입 16 LDA	58.4	10.6	PLP, dPLP, ddPLP 각 13 차에서 16 차로
특징벡터 차원증가	62.6	10.2	16 차에서 20 차로
음소당 데이터량에 따른 코드북 크기 지정	65.0	6.4	음소당 32-72 개 코드북
작업영역 변환 및 훈련	70.1	13.3	여행계획 데이터 5.4 시간분(이하 5,500 어휘대상)
Stack decoder 구현	72.2	6.7	
광역 음운환경 음소모델, 트라이그램 언어모델 도입	74.6	8.6	Tree lexicon, 변이음 자동 군집화 포함
데이터 증강	76.6	7.8	시간약속 10.2 시간, 여행 계획 14.9 시간
단기 침이모델도입	78.6	8.4	30msec 에서 10msec 가능
훈련반복	79.6	4.7	
UVS 정보활용	81.4	8.8	
화자적응	82.3	4.8	

위치에 항상 세심한 주의를 기울여야 한다. 따라서, 일반인들이 실용적으로 사용하기에는 여러 가지 면에서 불편함을 느낀다. 본 시스템에서는 이러한 문제점을 해소하기 위하여 원격음성입력기를 개발하였다. 즉, 여러 개의 마이크를 사용하여 음성신호를 받아들이고, 이들 음성신호간의 지연시간을 추정하여 동기시키고 더하므로써 신호대 잡음비를 향상시켰다. 평가 결과 40cm-80cm 의 거리에서 8 채널의 마이크 어레이를 사용할 경우 동일거리상의 1 개 마이크에 비하여 최대 5.9dB 의 신호 대 잡음비 개선을 할 수 있었으며, 16%의 인식오류 감소율을 얻을 수 있었다[2].

지식기반 탐색방법의 목적은 입력되는 음성신호에 대하여 명확하게 추출할 수 있는 정보를 추가로 활용하자는 것이다. 현재 HMM 기반의 음성인식시스템은 기본적으로 통계적 접근방식으로 음성신호에 대한 모든 정보를 통계적으로 처리한다. 그러나, 유무성음이나 묵음구간과 같이 비교적 검출이 쉬운 정보를 미리 추출하고 탐색시에 이들 정보를 기반으로 탐색공간의 절지(pruning)나 점수계산에 가산할 수 있다. 본 시스템에서는 재귀신경회로망을 도입하여 각 음성 프레임에서의 신호대 잡음비, 레벨교차율, 차분레벨교차율, 유성음 대역 에너지 비, 무성음대역 에너지 비 등의 정보를 활용하여 유성음, 무성음, 묵음을 식별하도록 훈련시키고 탐색에서 이 정보를 적용하였다. PLP 캡스트럼과 LDA 변환에 의한 음성특징을 사용하였을 때의 인식률은 79.6%를 나타내었으며 수프라세그먼트 정보를 같이 사용하였을 때는 81.4%로서 약 8.8%의 오류 감축률을 나타내었다[3]. 이 결과를 분석해보면 기존의 PLP 기반의 음성특징만을 사용하기 보다는 수프라세그먼트 정보를 추가함으로써 음성인식의 성능을 추가적으로 향상시킬 수 있음을 알 수 있으며, 앞으로는 음소군을 더욱 세분하여 시도할 예정이다.

음향모델의 규모를 설정하기 위하여 군집화 나무(clustering tree)의 분리에 사용되는 거리척도로서 기존의 엔트로피 차이와 함께 교차검증(cross validation)에 의한 교차 엔트로피(cross entropy)를 추가로 도입하였다. 즉, 훈련 데이타베이스를 여러 개의 부분으로 분리한 다음 각 부분만을 사용하여 인식기를 훈련한 확률분포를 각각 구한다. 그 다음 훈련 데이타베이스로부터 구한 분포간의 교차엔트로피를 구하여 이를 군집화 나무의 분리 조건에 사용되는 거리척도에 추가한다. 교차엔트로피는 두 확률분포간의 거리와 같은 개념으로서, 노드가 분리될수록 훈련 데이타베이스 조각에 의한 분포간에 차이가 커지므로 교차엔트로피도 커지게 된다. 따라서 엔트로피와 교차 엔트로피의 합이 0 보다 커야 한다는 조건을 분리의 종료 조건에 추가할 수 있다. 또한 노드가 분리될수록 그 노드에 할당되는 음성 샘플의 개수가 감소하여 분포의 강인성을 감소시키므로 이를 방지하는 장점도 있다. 실험결과 엔트로피 군집화에서 최대 인식률을 나타낸 코드북 개수 2,000 과 분포갯수 10,000 에 대하여 잎사귀 노드 개수 변화에 따른 인식률을 조사한 결과 엔트로피 기반의 인

식률과 차이는 크지 않았다. 이는 실험적으로 설정한 엔트로피 군집화와 결과와 잘 일치하는 것이며, 분포나무의 변경에 대하여 상당히 강인함을 보여주고 있다

한편, 인식의 단위로 어절을 사용하는 경우 처리 어휘수가 작업영역을 넓힘에 따라 급속하게 증가하여 확장성을 떨어뜨리는 점을 고려하여 언어학적 특성은 유지하되 소리값은 변하지 않는 의사형태소를 새로이 정의하여 수용을 시도하고 있다. 의사형태소는 주어진 어절의 소리값은 유지하는 범위 내에서의 언어학적인 형태소를 말한다. 즉 분리된 형태소들의 단순 결합(concatenation)에 의해서 원래의 소리값을 찾을 수 있음을 의미한다. 의사형태소는 일반적인 형태소와 매우 유사하나, 형태소의 분리에 있어서 소리값이 유지된다는 점이 매우 다르다[4].

그리고, 각 단어가 어떻게 소리나는가를 나타내는 발음열 생성기에 관하여 그간 단어내의 조음규칙만을 적용하여 왔으나 최근 문장내 끊어읽기 단위로 조음현상을 확장하였으며, 단어내 형태소 정보를 고려하여 조음규칙을 보완하였다. 또한, 음성인식 프로그램의 양이 방대하여 추가개발 및 관리가 어려워 짐에 따라 최근 객체지향 개념을 도입, C++ 언어를 사용하여 구현하므로써 소스프로그램의 양을 절반 이하로 줄였다[5]. 언어모델에서는 그간 사용하였던 바이그램(bigram)을 트라이그램 (trigram)으로 확장하였다.

4. 언어번역모듈

생략, 반복, 오발성, 간투사 등 대화체의 특성상 구문구조에 대한 정형문법을 사용하여 구문분석을 시도할 경우 대부분 실패하게 된다. 이에 정형문법 대신 발화의 의도 파악에 초점을 맞춘 개념기반의 문법을 작성하여 번역을 시도하고 있다[6][7].

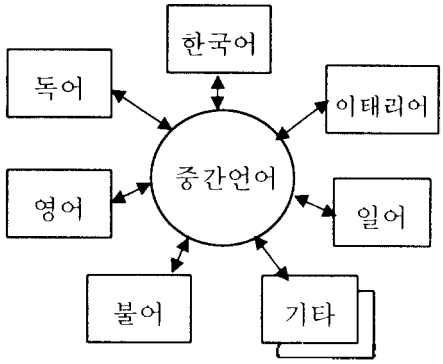

그림 3. 중간언어를 이용한 다국어간 번역

또한 모든 언어와 일일이 번역하는 대신 그림 3와 같이 다국어간 효율적인 번역을 위한 중간언어(IF: Interchange Format)를 C-STAR 회원기관과 함께 여행계획 작업 영역에 대하여 정의하고, 우리말과 중간언어 간의 번역기를 개발하고 있다[8].

중간언어의 구조는 화행(speech act)과 개념(concept), 그리고 변수(argument)와 그 변수에 해당하는 실제 값(value)로 구성된다. 화행이란 정보요청이나, 정보제공, 또는 인사 등 전체 문장의 큰 분류를 나타내는 것을 뜻하며, 개념은 보다 세부적인 정보 즉, 제기된 발화의 내용의 초점을 나타내는 것으로 객실 이용의 가능성 여부나, 여행상품, 비행기 등에 대해 정보 제공 등의 보다 명확한 의미를 나타내는 것들이다. 변수는 선택사항으로서 개념의 종류에 따라 구체적인 내용이 필요할 경우에만 존재한다. 그림 4에 중간언어를 통한 한국어와 영어간 번역의 일례가 예시되어 있다. 여기서 화행은 request-information, 개념은 availability-flight, 그리고, 개념을 보강하는 변수 destination과 via 및 해당 값으로 중간언어가 이루어져 있다. 한편, 우리말과 구조가 유사한 일본어와의 번역은 구문구조의 유사성을 활용하는 번역프로그램을 개발하였으며 그 적용성을 검토하고 있다.

그림 4. 중간언어를 이용한 한영 번역의 예

5. 음성합성모듈

합성음의 명료도와 자연성을 제고하기 위하여 여러가지 시도를 하여 왔다. 일찍이 TD-PSOLA 방식을 적용하여 반음절 단위 합성기를 구현하였고, 보다 긴 합성음편을 사용하여 연결점의

수를 줄여 왜곡이 일어날 가능성을 줄임으로써 명료도를 향상시킬 목적으로 트라이폰단위 합성기, 음절단위 합성기를 구현하였다. 이러한 시도는 당초 목적대로 보다 명료한 합성음을 얻을 수 있었으나, 한편으로 전후의 음운환경을 고려해야 하므로 보다 방대한 음성을 녹음할 필요가 생겼으며, 따라서 이에 대한 합성단위의 분절표기 및 피치 마킹 등의 작업에 많은 인력과 시간이 소요되게 되었다. 최근 이러한 문제점을 해소하기 위하여 자동훈련형 음성합성기를 개발하고 있다[9]. 즉, 음성인식모듈의 성능 향상에 따라 이를 활용, 합성단위 분절표기를 자동화하고, 녹음시에 래링고그래피(laryngography)를 사용하여 피치정보를 자동으로 추출하므로써, 합성데이타베이스 구축에 필요한 소요비용 및 시간을 대폭 절감시켰다. 기존의 트라이폰단위 DB의 경우 한 사람의 음성을 새로 수용하는데 5개월이상 소요되었으나, 자동훈련용 합성기의 경우 2주일안에 녹음과 DB구축을 끝내고 시작품의 음성을 들을 수 있었다. 자동 분절표기의 오류에서 올 수 있는 음질의 저하는 한 합성단위에 대하여 복수개의 음편을 유지하고 합성음 생성시 주변의 여건을 고려하여 최적의 음편을 선정하고 또한, 최장일치 선택방법 등의 음편의 최적접합점에 대한 척도를 도입하여 결정하므로써 방지하고 있다.

한편, 다층 신경망을 이용한 학습방법을 이용하여 음절의 시작, 가운데, 끝점에서의 에너지 크기를 결정하는 알고리즘을 구현하였으며[10], 합성음편 연결방법으로 한 피치 구간에서 음성의 대부분의 에너지와 고유 특성을 포함하고 있는 성문 닫힘 구간의 신호를 이용하여 성문 열림 구간에서의 음원신호를 추출 창함수를 적용하므로써 연결점에서의 신호왜곡을 최소화하는 알고리즘을 고안하였다[11]. 한편, 운율처리기술의 기반이 되는 K-ToBI 표준안을 작성하였다[12].

6. 결언

현재까지의 개발은 주로 각 모듈의 성능 향상에 초점에 맞추어져 왔으며, 앞으로는 성능향상 노력과 함께 위의 각 모듈을 통합하여 전체 음성언어번역 시스템을 구성할 예정이다. 상대 시스템과의 통신은 인터네트를 활용한 계획이며, 아바타를 도입하여 화상전송에 필요한 통신량을 최소화 할 예정이다. 또한, 음성언어번역 시스템의 실시간 동작을 위하여 입력 음성신호의 파이프라이닝 처리, 탐색 모듈에서 선행탐색(look-ahead)기법 등을 도입하는 방법이 검토되고 있다. 특히 보다 강력한 언어모델의 구축이 필요한 것으로 판단하고 있으며 현재 인식단위로 시도하고 있는 의사형태소가 유지하고 있는 언어정보가 이러한 작업에 도움을 줄 것으로 기대하고 있다.

참고문헌

[1] H.S. Lee, J. Park, and H.-R. Kim, "An Implementation of Korean Spontaneous Speech Recognition System," 1996 International Conference on Signal Processing Applications & Technology, pp. 1801-1805, Boston, USA, Oct., 1996.

[2] Youngjoo Suh, Jun Park, and Youngjik Lee, "A user friendly remote speech input unit in spontaneous speech translation system," ESCA-NATO Tutorial and Workshop on Robust Speech Recognition for Unknown Communication Channels, 1997.

[3] Youngjoo Suh, Kyuwoong Hwang, Oh-Wook Kwon, and Jun Park, "Utilizing the voiced, unvoiced, and silence information to improve the performance of speech recognizer," ITC-CSCC 98, pp. 669-672, 1998.

[4] 권오욱, 박준, 황규웅, "의사형태소 단위 대어휘 연속 음성 인식기 개발," 제15회 음성통신 및 신호처리 워크샵, 발표예정, 1998.

[5] 황규웅, 권오욱, 박준, 서영주, "C++언어와 Standard Library를 이용한 음성인식기 개발," 제15회 음성통신 및 신호처리 워크샵, 발표예정, 1998.

[6] 최운천, 한남용, 김재훈, "대화체 음성언어번역 시스템에서의 개념기반 번역 시스템," 한국정보처리학회 논문지, 제4권 제8호, 1997년 8월, pp.2025-2028.

[7] N. Han, U. Choi, and Y. Lee, "An Implementation of Partial Parser in the Spoken Language Translator," ICASSP '98, pp. 205-208, Seattle, 1998.

[8] 최운천, "다국어 대화체 음성언어번역 시스템을 위한 IF와 IF 태깅," 제15회 음성통신 및 신호처리 워크샵, 발표예정, 1998.

[9] 김상훈, 이정철, 강동규, 이영직, "대용량 운율 데이터를 이용한 자동 합성 방법," 제15회 음성통신 및 신호처리 워크샵,

발표예정, 1998.

[10] Jungchul Lee, Donggyu Kang, Sanghun Kim, and Koeng-mo Sung, "Energy Contour Generation for a Sentence Using a Neural Network Learning Method," To be published in ICSLP 1998.

[11] 강동규, 김상훈, 이정철, "성분 단위 구간 가변에 의한 피치 변경," 제15회 음성통신 및 신호처리 워크샵, 발표예정, 1998.

[12] Sanghun Kim, Jungchul Lee, and Jun Park, "Standardization of Korean ToBI system and Autolabeling Using Low-High Intonation Stylization," Proc. of ICSP '97, Seoul, Korea, 1997.