제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

대용량 운율 음성데이타를 이용한 자동합성방식

김상훈, 이정철, 강동규, 이영직
한국전자통신연구원, 휴먼인터페이스연구부, 음성신호처리연구팀

Automatic Synthesis Method Using Prosody-Rich Database

Sanghun Kim, Jung-Chul Lee, Dong-Gyu kang, Youngjik Lee
Speech signal processing team, Human interface department, ETRI
ksh@zenith.etri.re.kr

Abstract

In general, the synthesis unit database was constructed by recording isolated word. In that case, each boundary of word has typical prosodic patterns like a falling intonation or preboundary lengthening. To get natural synthetic speech using these kinds of database, we must artificially distort original speech. However, that artificial process rather resulted in unnatural, unintelligible synthetic speech due to the excessive prosodic modification on speech signal. To overcome these problems, we gathered thousands of sentences for synthesis database. To make a phone level synthesis unit, we trained speech recognizer with the recorded speech, and then segmented phone boundaries automatically. In addition, we used Laryngo graph for the epoch detection. From the automatically generated synthesis database, we chose the best phone and directly concatenated it without any prosody processing. To select the best phone among multiple phone candidates, we used prosodic information such as break strength of word boundaries, phonetic contexts, cepstrum, pitch, energy, and phone duration. From the pilot test, we obtained some positive results.

1. 연구배경

최근 5~6 년간의 합성분야의 연구동향은 대량의 합성 데이터베이스로부터 합성단위를 인결하여 합성하는 concatenation synthesis 방식이 합성연구의 주류를 이루고 있으며, 특히 음성파형에서 운율조절이 가능한 TD-PSOLA 방식이 제안됨으로써 명료도가 높은 합성음을 생성할 수 있었다. ETRI 에서는 1992~1993 년, 국내에서는 처음으로 TD-PSOLA 방식을 채택, 합성단위인 CDU(Context Dependent Units)를 고안하여 기존 파라미터(LPC 계열) 합성방식보다 훨씬 명료한 무제한 낭독체 남/여성 합성기를 개발하였고, 1994~1996 년에는 남/여 합성음의 자연성 개선을 위한 운율 처리 모델을 개발하였다. 1997 년에는 반음절 접합으로 인한 연결집점에서의 왜곡을 줄이기 위해 주변 음운환경을 고려한 음절단위인 CDS (Context Dependent Syllable)를 제작하였고 이를 합성기에 채택하였다. 또한 트라이폰(Triphone)단위를 사용한 합성기도 개발하였다. 이 밖에 LG, 삼성에서 상용화된 합성기를 개발하여 합성기의 저변화에 기여한 바 있다. 외국의 경우 대용량 음성 데이터베이스를 기반으로 하는 합성시스템에 관한 최근 연구 동향을 요약하면 표 1 과 같다[1].

표 1: Recent research trend

Author	DB	Unit	Signal pro.	Evaluation
Hauptmann (1993)	*360MB*	*Phones*	*PSOLA*	*Nearly indistinguishable barely intelligible*
Campbell (1995)	*1 hour*	*Phone segments (CHATR)*	*No or PSOLA*	*Very natural*
Nakajima (1993)	*627units*	*Phone (COC)*	*LPC*	*intelligible*
Sagisaka (1988)	*5240 words*	*Non uniform*	*LPC*	*?*
Donovan (1995)	*1 hour*	*HMM-state sized segments*	*LPC or PSOLA*	*natural*
X.Huang (1996)	*6,000 sentences*	*" (WHISTLER)*	*LPC*	*natural*

1988년 NTT의 Nakajima는 음성신호에 기반을 둔 bottom-up 방식인 Context Oriented Clustering을 이용하여 합성단위를 자동 생성하였다[2]. 고려한 환경변수는 음운환경, 3 stress levels, word final position, sentence final position이며, 5분 정도의 음성데이터로 627 개의 합성단위를 생성하여, 연결구간에서 LPC 파라미터의 스무딩(smoothing) 과정없이 매우 명료한 합성음을 생성하였다. 1994년 NTT의 Itoh는 "ML-COC+PSOLA", 즉 multi-lingual COC를 제안하여 영어에서도 COC 기법을 적용할 수 있음을 보여주었다. 1988년 ATR의 Sagisaka는 음소단위로 최장일치되는 합성단위를 제안하였는데, 이 단위를 이용함으로써 합성단위간 연결부 개수를 줄일 수 있었으며, 일본어의 경우, 3~4개 음소열 중 약 20%로도 무제한 텍스트에서 80%를 커버한다고 한다[4]. 그러나 최장일치 방법은 합성단위의 개수가 크게 증가되는 단점이 있다. 이에 최근에는 합성단위의 개수를 줄이기 위한 시도로서 빈번히 발생하는 최장 단위를 찾고 이를 저장, 사용하는 방식으로 전환하고 있다. 빈도를 고려한 연구로는 1994년 Klavans은 트라이폰의 발생빈도를 조사한 바 있다. 최장일치를 사용한 합성기에 대한 음질에 대해서는 아직 알려지지 않은 것 같다. 다만 5,240개의 어절과 LP synthesizer를 이용하여 합성하였다고 한다.

1993 년 Hauptmann 은 3,253 문장(360MB)을 음성인식기를 이용하여 115,000 개의 음소로 분절하고, 강세정도,

유운환경, 음절, 어절, 문장내 위치를 고려하여 음소를 선정, 합성하는 방식을 제안하였다. 분절시 음성인식기에는 음소열이 주어지며, 선정된 음소는 PSOLA 합성방식을 이용하여 연결된다.

1995 년 ATR 의 Black 과 Campbell 은 Hauptmann 방법을 좀 더 최적화한 방식을 이용한다. 즉 최적 음소는 연결구간에서의 켑스트랄 거리(cepstral distance)에 해당하는 continuity cost function 과 운율환경이 최적화되는 target cost function 을 최소화하는 단위로 선정된다. Continuity cost 는 음성적, 운율적 요소가 자연스럽게 연결되도록 하며, target cost 는 규칙에 의해 생성된 운율에 가장 가까운 음소를 음운환경, 지속시간, 에너지, 평균 피치를 이용하여 선택하도록 한다. Target cost 가 최소화될 경우, 운율처리를 위한 신호처리 과정에서의 신호왜곡을 최소화할 수 있다. 특히 음소간 연결시, 연결구간 7 프레임 동안 VQ index 간 거리가 가장 작은 연결지점을 선정하므로 음성인식기의 자동 분절 오류에 강인하다고 할 수 있다. 그러나 이 방법은 합성음의 음질이 문장의 일부분에서 매우 자연스럽기나 또는 다른 부분에서 매우 부자연스러운 합성음을 생성해내는 단점이 있어 대용량의 DB 확보가 필수적인 방법이 된다[3].

1995 년 Cambridge Univ.의 Donovan 는 인식기의 향상된 성능을 합성에 이용하고자 HMM trainable synthesizer 를 구현하였다. 이 시스템은 decision tree state clustered HMM 을 이용, 유사한 state 를 그룹화하며, HMM-state sized segment 단위를 연결함으로써 합성을 한다. HMM 훈련에 사용된 음성데이터는 약 1 시간 분량의 연속음이며 데이터가 부족할 경우, backing off, decision trees, parameter smoothing, bottom-up clustering 방법을 이용하여 음운환경을 줄인다. 합성 DB 구축에서부터 합성음까지는 모두 자동으로 이루어지므로 다른 화자, 다른 언어로의 전환이 매우 용이한 방식이다.

1996~1997 년 Microsoft 사에서는 실제 상용화된 HMM trainable synthesizer 로서 "WHISTLER" 라는 합성기를 개발하였다. 이 합성기는 약 6,000 문장의 훈련 DB 를 이용하며, HMM 의 senone 단위를 합성단위로 한다. 합성방식으로는 LP synthesizer 를 이용하고 있으며, 각 state 에 있는 복수개의 LP parameter 중 중심 벡터를 사용하며, 자연스럽고 명료한 합성음을 생성해낸다[5].

2. 연구동기

음성파형을 그대로 이용하는 TD-PSOLA 방식에서는 음소경계, 피치 등의 정보가 매우 정밀하게 분절되어야 하기 때문에 합성 데이터베이스를 제작하는데 상당한 기간(보통 3~5 개월)이 소요된다. 더구나 합성 데이터베이스가 완성되었다 하더라도 최종 합성음의 음질은 보장하지 못한다. 특히 국외 연구동향으로부터 알 수 있듯이 하드웨어의 기술적 발전이 합성 DB 크기의 제약을 점점 줄이고 있기 때문에 합성단위의 특성도 음운환경 뿐 아니라 피치, 지속시간, 에너지 등 운율적 요소까지 고려된 합성단위를 사용하려는 시도가 이루어지고 있다[3]. 이에 따라 음성 DB 의 크기는 2~3 년 전에 비해 약 20 배 이상 증가되었고, 몇 사람의 전문가가 과거 수동으로 해오던 합성단위의 분절 과정으로는 약 4~5 개월의 기간이 소요된다[6][7].

표 2: 합성기의 productivity & quality 비교

	ETRI			CHATR	whistler
단위	CDU	반음절	CDS	Phone	Subphone
DB 크기	15MB	15MB	120MB	1 hour	6,000 문장
녹음	3 시간	3 시간	1.5 개월	2-3 hour	?
분절	1 개월	2 주	2 개월 (반자동)	자동	자동
피치	1 개월	1 주	1 개월	No	laryngo
Total time	2 개월	0.7 개월	3 개월	?	2 일
명료도		2.8	3.0	3.5	3.5(영)
자연성		2.9	2.9	3.2	3.1(영)

따라서 음성 DB 의 구축에 자동화 과정이 절실히 필요하다. 이미 타 기관에서는(Cambridge Univ., Microsoft Co., ATR) 성능이 향상된 음성인식기를 이용하여(표 3) 합성단위를 자동생성하고 있으며 음성인식에서 사용하는 알고리즘인 Viterbi search, decision tree 등을 적용하여 합성단위간 연결왜곡을 최소화하고 있다. 이에 따라 ETRI 에서는 과거 합성기 개발 경험과 이에 따른 문제점을 분석하고 표 2 와 같이 타 기관의 생산성(productivity) 및 합성음질(quality)에 측면에서 비교하여 글소리 IV 버전 ETRI 합성기 개발에 반영하고자 한다.

표 3: 음소 분절 성능

	human	automatic
Continuous speech	8msec , 80%	11.5msec , 80%
Isolated word	95%, 30msec(1991)	89.5%, 30msec(1993)

당 연구실에서 1992~1997 년 동안 합성단위 구축에 어려웠던 문제점과 이 문제점에 대해 타 연구기관에서 대처하는 방식은 다음과 같다.

- 발성의 어려움

CDU 합성단위는 대부분 2 음절로 이루어진 단어이며, 무제한 합성용 DB 를 제작하기 위해 발성이 어려운 무의미어 단어로 이루어져 있다. 따라서 발성자는 그 단어에 익숙치 않아 부자연스러운 발성을 하게 된다. 특히 발음에 충실하기 위해서 과도하게 조음이 되는 경우도 있어 오히려 문장합성음에서는 부자연스러운 요인으로 작용한다. CDS 합성단위용 발성 리스트는 의미가 있는 단어로 이루어져 있으나, 단어내 형태소 경계가 올 경우 음운 환경에 부자연스러운 발성을 하게 된다. CHATR 나 whistler 인 경우, 주로 문장 단위로 발성하게 되어 자연스러운 발성이 이루어진다.

- 음소단위 분절 및 피치 검출

이 문제는 합성 DB 가 대용량화 됨에 따라 합성단위, 합성방식에 관계없이 발생한다. CHATR 나 whistler 의 경우, 음성인식기를 이용하여 DB 를 자동 분절하며, 피치 추출도 래링고그라프를 이용하므로 합성 DB 구축에 소요되는 시간을 최소화하고 있다.

- 합성단위 연결점의 최적화

CDU 는 전문가에 의해 음소 분절되어 접합점의 위치가 고정식으로 정해져 있고 단일후보만 사용하므로 접합점에서의 불일치(스펙트럼, 피치)가 큰 경우 이를

피할 수 없다. CHATR 의 경우, 합성단위를 복수개 후보로 하고 이중 접합점에서의 왜곡이 가장 작은 후보를 선정하여 불일치를 최소화한다.

- 과도한 운율조절
어절발성은 일반적으로 어절끝 음절에서 하강하는 억양과 경계전 지속시간이 길어진다. 그리고 어절내 음소의 길이가 문장내 음소길이보다 전반적으로 길어진다. 따라서 문장운율을 적용할 경우, 과도한 운율조절(억양의 상승, 지속시간 단축)이 불가피하며 합성음질 저하의 원인이 된다.

- 음운환경의 제약
CDU 의 경우, 한국어의 주요 음운환경만 고려되어 있어 합성음의 명료도, 자연성에 한계가 있다. 따라서 트라이폰이나 CDS 와 같이 최대한 많은 음운환경을 고려하되, 합성 DB 의 크기와 합성단위의 효율성을 고려하여 유사 음운환경을 사용하거나 음운환경의 그룹화(context clustering) 방법이 필요하다. Whistler 의 경우 decision tree 를 이용하여 음운환경을 몇 개의 그룹으로 나눈다.

- 수작업에 의한 일관성 결여
수작업에 의한 분절은 자동 분절에 비해 정확하다고 할 수 있으나, 여러 사람이 작업을 하거나 데이터가 방대해질 경우 일관성이 떨어진다. 물론 수작업 하기 전 분절 기준이 확실히 설정되어 있다면 이 문제는 다소 줄어들 수 있다. 자동 분절은 일관성은 있지만 정확도에 있어 수동보다 성능이 떨어진다. CHATR 의 경우, 합성단위 연결부의 7 프레임동안 VQ index 간 거리가 가장 작은 연결점을 찾기 때문에 자동 분절의 오류를 줄일 수 있다.

이에 따라 음성인식기, 래링고그라프를 사용하여 합성 데이터베이스 제작에 걸리는 시간을 최소화하고, 과도한 운율조절로 인한 명료도, 자연성의 저하를 막기위해 다양한 운율현상이 포함된 문장단위를 녹음하여 합성단위를 추출하는 방식으로 합성기를 개발하고자 한다.

3. 시스템 개요

이 시스템은, 합성단위로는 복수개의 트라이폰을 사용하며, 최적 트라이폰은 운율 파라미터와 어절 경계 정보를 이용하여 선정하며 운율조절 없이 직접 음성파형이 연결되어 합성된다.

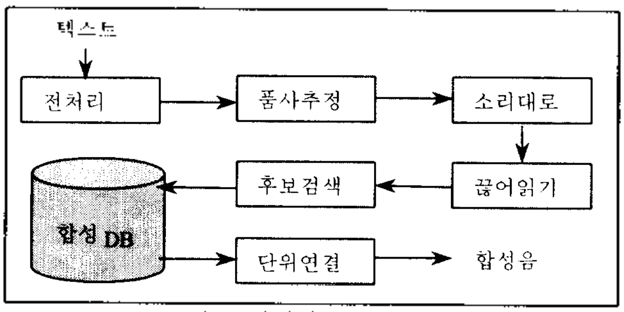

그림 1: 합성시스템 개요도

현재 사용중인 문장 수는 2,092 문장이다. 끊어읽기 추정은 bigram 을 사용하여 30 msec 이상의 휴지길이를 갖는 위치를 예측한다. 구축된 합성 데이터베이스의 크기는 음성, 운율파라미터, 피치 등을 포함하여 약 544MB 이다.(표 4)

표 4: 합성 데이터베이스의 크기

기능	크기(Byte)
DBM 파일	4,096
	920,576
매핑 테이블	78,496
피치 데이터	9,328,320
음성 신호, 묵음 포함 16 kHz, 16 bit	497,182,700
켑스트럼 / 운율 데이터	37,006,624
Total size	544,520,812

3.1 발성문장 리스트 추출

합성 단위로 사용하는 트라이폰은 음성인식에서 사용하는 단위와 동일하다. 즉 음소를 기준으로 좌우 음운환경이 다르면 하나의 트라이폰이 된다. 분절은 음소경계를 찾게 되며, 합성단위의 연결도 음소경계에서 연결되도록 한다. 트라이폰을 음성합성 단위로 할 경우 약 25,000 개의 합성단위가 필요하나 이를 합성 DB 에 모두 등록할 수 없기 때문에 고빈도로 발생하는 트라이폰을 우선적으로 등록하는 것이 효율적이다. 고빈도 트라이폰을 포함하는 최소 문장 set 을 선정하기 위해 다음과 같은 조건을 고려한다.

- 음운환경이 일치해야 한다.
- 고빈도 트라이폰을 한 문장내 최대한 많이 포함하여야 한다.
- 선정된 문장 set 은 트라이폰 빈도 커버리지가 최대가 되도록 해야 한다.
- 선정된 문장개수가 최소화되어야 한다.

문장 set 은 다양한 장르에서 무작위로 추출한 1,000 문장과 텍스트 코퍼스 2 만 문장에서 위의 조건을 고려한 후 최적화 알고리즘을 사용하여 1,092 문장을 추출하였다. 최종적으로 합성 DB 로 사용되는 2,092 문장에 포함된 유일한 트라이폰 개수는 12,000 여 개이며, 이 트라이폰들은 297 만 어절 텍스트 코퍼스를 이용하여 트라이폰 빈도 커버리지를 구했을 때 약 99.5%를 커버할 수 있다. 다시 말해서 12,000 개의 트라이폰으로 약 297 만 어절에서 발생하는 트라이폰의 빈도를 99.5% 커버할 수 있음을 의미한다. 참고로 297 만 어절로부터 발생하는 유일한 트라이폰수는 19,198 개이며, 총 트라이폰수는 19,391,918 개가 발생했다.

3.2 전사

자동분절을 하기위해 발성한 음성과 발음이 일치하는 텍스트가 필요하다. 따라서 이 과정은 먼저 발음변환규칙을 적용하여 소리나는데로 바꾼 다음 청각에 의존하여 수정, 작성된다. 현재 청각에 의해 확연히 구별되는 오류를 주로 수정한다. 수정내용은 다음과 같다.

- 발성이 텍스트와 다른 경우 수정
- 발음변환 규칙 오류 수정
- 예외 발성(특히 경음화)을 수정한다.
- 복합어의 어절 경계를 재설정한다.
- 띄어쓰기는 맞춤법을 따른다.

그리고 현재는 고려하고 있지않으나 전사할 때 정해야 할 기준이나 문제점을 보면 다음과 같다.

- 현재의 음성인식기는 초성, 종성을 구별하지 않는다. 따라서 어절(또는 형태소단위)간 연음이 될 경우, 연음된 종성자음이 초성으로 발성되었는지, 아니면 종성으로 발성되었는지 구별할 수 없다.(예: 아버지는 아이가 (분절전)-> 아버지느 나이가(분절 후))
- 발음이 애매한 경우 전사 방법
- /ㅎ/ 탈락에 의한 발음 표기
- 화자습관에 따른 발음
- 부정확한 발성: 더듬거림, 조음 부정확

3.3 자동 분절 및 피치 추출

음성합성 단위인 음소를 생성하기 위해 자동 분절을 수행한다. ETRI 음성인식시스템은 FM radio news 문장, 대화체 문장 및 낭독체 문장 등에서 분절 대상 음소의 약 80% 이상이 오류가 30msec 이내인 범위로 분절되며, 고립단어에 대해서는 약 60%의 성능을 보여주고 있다 [8]. 음소분절에 사용되는 음성인식기는 다수 화자에 의해 훈련되어진 파라미터를 사용함으로 새로 추가되는 화자의 음소경계 추정시 다소 일관성이 떨어질 수 있다. 따라서 새로운 화자에 대해 재훈련을 수행하여 음소경계 추정의 일관성을 높인다. 그림 2는 음성인식기가 재훈련 되었을 때 중심벡터의 이동을 보여주며 이로부터 새로운 화자의 음소분절의 일관성을 높여주는 효과가 있다. 비록 자동 분절결과가 수동분절결과와 다르다 할지라도 일관성이 유지된다면 합성단위간 연결점에서의 왜곡은 최소화 될 수 있다. 그림 3은 자동 분절 결과 및 피치를 추출하기 위한 래링고그라프의 출력파형을 나타내고 있다.

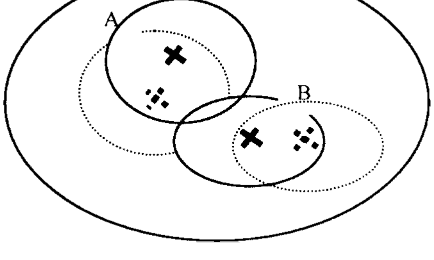

그림 2: 음소 A,B 의 훈련 파라미터 분포 예측
(훈련전: 점선, 훈련후: 실선, X:중심벡터)

이 시스템은 현재 TD-PSOLA 방식을 사용하고 있지 않기 때문에 정교한 피치 추출은 필요하지 않으나 래링고로부터 추출된 피치는 거의 정확한 피치값을 추출할 수 있으며, 추출된 피치값은 여러 운율 파라미터의 값 추출, 자동 분절의 후처리 등에 사용되고 있다. 또한 향후 운율모델 개발, TD-PSOLA 방식의 적용 등을 위해 합성 DB 에 피치값이 포함되어 있다.

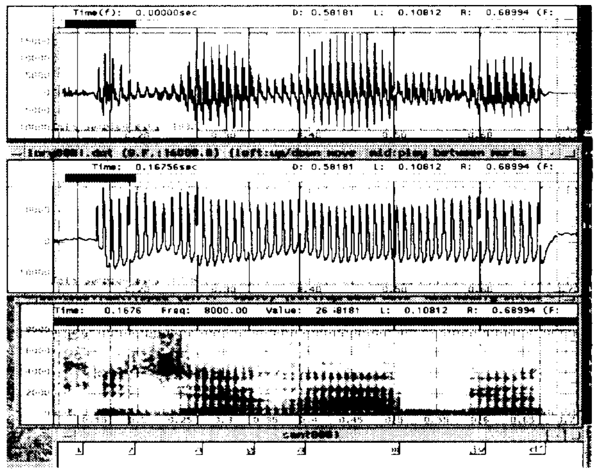

그림 3: 자동 분절결과
(a.음성파형, b. 래링고 결과, c. 스펙트럼 및 분절결과)

3.5 래링고 신호를 이용한 분절 위치 조정

ETRI 음성인식기는 일반적으로 그림 4 와 같이 유성음과 무성초성자음 환경에서, 무성초성자음의 분절이 유성음부쪽으로 치우쳐서 분절된다. 이 분절정보를 이용하여 음성을 직접 연결할 경우, 무성음부에 포함된 유성음에 의해 합성음의 음질을 저하시키게 되는데 이를 줄이기 위해 래링고그라프 신호를 이용하여 최대한 유성음이 무성음에 포함되지 않도록 자동 분절위치를 조정한다.

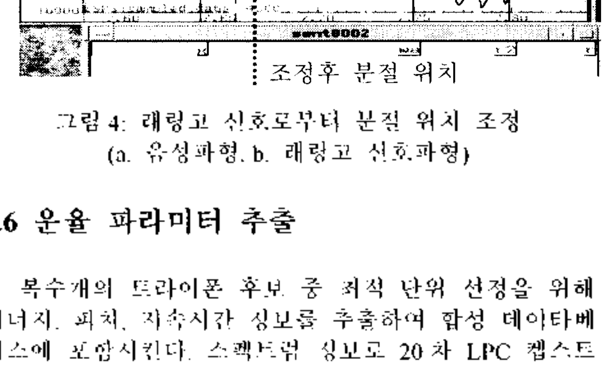

그림 4: 래링고 신호로부터 분절 위치 조정
(a. 음성파형, b. 래링고 신호파형)

3.6 운율 파라미터 추출

복수개의 트라이폰 후보 중 최적 단위 선정을 위해 에너지, 피치, 지속시간 정보를 추출하여 합성 데이타베이스에 포함시킨다. 스펙트럼 정보로 20 차 LPC 켑스트럼을 사용하였고, 에너지는 데시벨로, 피치는 Hz, 지속시간은 msec 단위로 변환하여 저장한다.

다양한 운율현상을 가진 문장 단위에서 합성단위를

추출하므로 합성단위 각각은 문장에서의 운율정보를 가지고 있어야 한다. 이를위해 우선 트라이폰 단위로 최장일치를 보장하기 위해 음소 주변의 스펙트럼 정보를 이용, 최장 트라이폰 열이 합성단위가 되도록 DB를 구성하였다. 한 어절이 트라이폰 A, B, C의 열을 포함할 때 트라이폰 B의 경우, 좌측 음소(A)의 경계에 해당하는 1 프레임(300 샘플)에 대한 켑스트럼 값과 A 음소 경계에서의 피치값, 그 피치의 에너지, 그리고 음소 A의 지속시간으로 좌측 음운환경을 저장하며, 우측 음운환경에는 현재 음소 즉, B의 경계에 해당하는 켑스트럼 값과 B 음소 경계에서의 피치값, 그 피치의 에너지, 그리고 음소 B의 지속시간이 저장된다. 각 트라이폰은 어절내 인접 트라이폰의 켑스트럼 값을 가지고 있으므로 무작위로 구성된 트라이폰 합성 DB로부터 최장일치가 되게 트라이폰을 선정할 수 있게 된다. 또한 에너지, 피치, 지속시간 정보를 이용하면 어절 합성할 때 음소간 운율변화를 고려할 수 있는 트라이폰을 선정할 수 있다.

3.7 어절 경계 강도 할당

어절 경계 강도는(Break indices) 화자의 자연스런 발성에 따라 형성되는 것으로 지속시간, 억양, 휴지 등의 음향적 변화로 경계지어지며, 적절한 어절 경계 강도 할당은 음성합성의 자연성을 크게 향상시킬 수 있다[9]. 일반적으로 경계현상은, 통사단위 혹은 의미단위의 경계가 주어진 환경에 따라 운율적인 경계로 실현될 때, 그리고 생리학적인 숨쉬기의 한계 및 호흡과 관련지어 나타나는 지각적인 끊김의 현상을 말한다. 경계의 운율 현상은 휴지구간 뿐 아니라 휴지구간이 없이 마지막 음절의 장음화만으로 이루어지기도 하고, 억양의 변화로 나타나기도 하며 음질(voice quality)의 변화로 나타나기도 한다. 이와 같이 어절경계강도를 결정짓는 다양한 운율현상이 있으나 현재 이 논문에서는 어절의 경계강도를 자동 레이블링 결과로 알 수 있는 휴지길이 정보만 이용한다. 그리고 경계 유형은 경계가 없는 강도(zero), 약한 경계 강도(minor break strength), 강한 경계 강도(major break strength) 3 단계로 정한다[10].

끊어읽기 유형은 같은 문장이라도 사람에 따라 다를 수 있고, 또한 한 문장에서도 다양한 끊어읽기 패턴이 올 수 있다. 그러나 주요하게 끊어지거나 기의 끊어지지 않는 어절경계인 경우는 추정시 반드시 지켜지는 것이 합성음의 자연성에 중요하다. 다행히 이러한 경계 유형은 일반적으로 확률치가 매우 높거나 매우 낮아 텍스트상에서 경계예측방법으로 bigram과 같은 간단한 방법을 사용해도 된다.

Bigram은 어절간 긴밀도를 알아보는 1차적인 척도이며, 빈번히 강한 경계가 오거나 거의 경계가 오지 않는 규칙은 어느 정도 반영할 수 있다. Bigram 확률은 2,092 문장의 음소분할 데이터로부터 추출할 수 있다. 3 단계 경계강도는 휴지길이 30msec를 기준으로 설정하였으며, 이 휴지길이는 다르게 설정될 수도 있다. 이렇게 합성 DB와 운율 DB가 공동적으로 사용되므로 화자의 끊어읽기 특성을 자연스럽게 규칙에 반영할 수 있다. Bigram의 성능은 2가지 텍스트 데이터를 이용한 open test에서 각각 46.0%, 38.2%의 강한 경계 강도 예측정확률과 22.8%, 8.4%의 삽입오류율의 성능을 내었으며, 참고로 trigram인 경우, 58.3%, 42.8%의 강한 경계 강도 예측정확률과 30.0%, 11.8%의 삽입오류율을 나타낸다[11]. 그러나 이 합성시스템에서는 합성단위 선정 등 끊어읽기가 매우 중요하므로 좀 더 정교한 알고리즘이 요구된다.

3.9 최적합성단위 선정 및 합성

임의의 어절을 합성하기 위해 이를 트라이폰열로 변환하고, 각 트라이폰의 복수후보를 합성 DB에서 가져온다. 트라이폰당 평균 n개 정도의 복수후보가 있다면 상태(state)간 약 n^2개의 경로(path)가 생기게 된다. 이들 경로로부터 복수개의 트라이폰중 가장 왜곡이 적은 경로를 찾기 위해 비터비탐색(또는 dynamic programming)를 수행한다. 왜곡은 식 (1)와 같이 각 상태간 유클리디언 거리를 사용하여 최종 상태까지 누적한다. 각 특징벡터간 가중치를 가하고, 가중치는 지각 실험에 의해 "trial and error"로 결정한다.

$$Distance_{path} = \sum_{i=1}^{STATE} [w_{pitch}(Pitch_i - Pitch_{i-1})^2 + w_{power}(Power_i - Power_{i-1})^2 + w_{dur}(Dur_i - Dur_{i-1})^2 + w_{cep}(Cep_i - Cep_{i-1})^2] \quad \text{.......(1)}$$

$$\text{Optimal path} = Min_{path=0,n}(Distance_{path}) \quad \text{.......(2)}$$

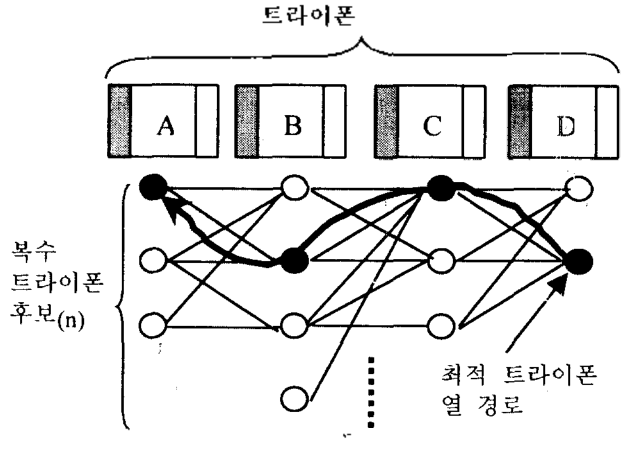

그림 5: 비터비 탐색수행도

합성하고자 하는 어절이 트라이폰 A,B,C,D로 이루어졌을때 그림 5에서와 같이 비터비탐색을 수행한다. 전방향 과정에서, 각 상태에서의 트라이폰 복수 후보들은 다음 상태의 트라이폰과 왜곡을 계산하게 되며 최종상태에서 역방향으로 최소 누적왜곡을 가진 최적경로를 찾는다. 이때 왜곡이 상태간 0(zero)일때 같은 어절에서 인접하여 발생하는 트라이폰임을 알 수 있는데 현재 구현된 방식에서는 식 (2)와 같이 어절내 트라이폰열의 누적 왜곡이 최소화되는 경로를 찾으므로 반드시 최장일치가 되는 단위를 사용하고 있는 것은 아니다. 참고로 같은 어절내에서 발생하는 트라이폰을 연속하여 선정되는 빈도를 조사하면 다음과 같다. 이 결과는 297만 어절에 포함되지 않은 164,460개 어절 중에서 유일한 어절을 추출하여 사용했으며, 유일한 어절 개수는 47,828개, 총 트라이폰 개수는 311,149개가 발생한다. 각 어절은 평균 6.51개의 트라이폰 열로 이루어져 있다.

비터비탐색을 이용하여 어절내 왜곡이 최소화되는 트라이폰열을 선정했을 때, 각 어절당 4.66 개의 3 음소열과 0.92 개의 4 음소열, 0.48 개의 5 음소열, 0.27 개의 6 음소열 등으로 최장일치된다. 음소열이 3 인 경우, 1 개의 트라이폰이 선정되며, 음소열이 4 개인때 2 개의 인접 트라이폰을 사용한다. 여기서 인접 트라이폰은 같은 어절에서 연이어 발생하는 트라이폰을 말한다.

3.10 Artificial process

합성 DB 에 등록되지 않은 트라이폰이 발생한 경우, 우선 음운환경이 유사한 트라이폰을 사용하며, 유사 음운환경이 없는 경우에는, 음운환경에 무관하게 "context independent phone"을 사용한다.

합성 DB 가 대용량화 됨에 따라 녹음된 아나운서의 음성도 시간에 따라 변해 합성음질이 불안정해진다. 특히 음색의 변화, 에너지의 변화 등이 합성음질을 저하시키는 큰 요인이 된다. 따라서 전체 음성에 대해 음색이 다른 발성은 합성 DB 구축할때 제외시키는 과정과 에너지를 일정한 크기로 정규화하는 과정이 필요하다. 이번 연구에서는 에너지를 정규화시키는 과정만 수행하였다. 에너지는 각 어절당 샘플의 피크치를 찾고 이를 미리 정한 레벨의(여기서는 25,000) 크기값과 비를 구한 후, 이 비율값을 어절내 매 샘플에 곱한다. 이와 같은 과정은 원음성의 자연스러운 에너지 컨투어에 왜곡을 가져올 수 있으나 전반적으로 이 과정을 수행하므로써 합성음이 안정화되었다.

4. 실험 결과

생성된 합성음은 명료하고 자연스러운 부분도 있으나 전반적으로 불안정하다. 즉 합성단위가 없는 경우 부분적으로 불명료하거나, 에너지, 피치 및 빠르기가 갑자기 변하는 곳도 발생해 부자연스러운 합성음을 생성한다. 그러나 문장단위로부터 합성단위를 추출했고, 과도한 운율조절이 없이 합성단위를 선정, 연결하므로 기존 합성음의 기계적인 소리와 다르게 사람이 말하는 것 같은 자연성은 얻을 수 있었다. 특히 합성데이타베이스 제작이 거의 자동으로 이루어지므로 새로운 화자에 대해서도 단시간내에 구축할 수 있다. 이번 연구를 통해 개발된 합성시스템과 기존의 합성방식과 다른점이 있다면 다음과 같다.

- 무의미 어절 발성의 어려움: 자연스러운 문장 단위(단위시간당 트라이폰 개수가 더 많음)
- 음소단위 분절 및 피치 검출에 많은 시간이 소요: 음성인식기(음소분할), 래링고(pitch 검출)로 자동화
- 합성단위 연결부에서 불연속성 발생: 복수후보를 사용
- 과도한 운율조절: 운율 조절하지 않음
- 음운환경의 제약: 트라이폰 사용 및 음운환경 그룹화
- 수작업에 의한 일관성 결여: 자동이므로 일관성 보장
- 생산성 향상: 1 개월 이내 새로운 화자에 대한 합성단위 데이터베이스 제작 가능, 따라서 다양한 화자에 대한 합성음을 단시간내에 확보할 수 있음.

5. Further works

현재 합성 DB 를 보강하고 있으며, 상용화를 위해 DB 압축 연구도 진행하고 있다. 그외 이 시스템의 문제점과 향후 연구되어야 할 내용은 다음과 같다[12].

- 부족한 트라이폰은 보완
- 음색이 현저히 다른 단위는 합성 DB 에서 제외
- 어절 경계에서의 억양 패턴만 적용: KToBI 의 상승, 하강, peak 값, 상대적 LH 비율 이용
- 구 단위 구성을 위해 Break indices 추출
- 자동분절 후처리
- 음소분할 오류를 고려한 단위 연결 방법

감사의 글

이 연구는 정보통신부 출연 "HCI 를 위한 음성입출력 처리기술 개발"과제의 연구 결과입니다.

참고문헌

[1] Nakajima S. and Hamada H., "Automatic generation of synthesis units based on context oriented clustering", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, New York, pp. 659-662, April 1998.

[2] Sagisaka Y., Kaiki N., Iwahashi N. and Mimura. K., "ATR v-Talk speech synthesis system", *International Conference on Spoken Language Systems*, Banff, Canada, pp. 483-486, 1992.

[3] Black, A.W., and Campbell, N., "Optimizing Selection of Units from Speech Databases for Concatenate Synthesis", *Proceedings of EUROSPEECH'95*, Spain, pp.573-576, 1995.

[4] Donovan R.E. and Woodland P.C., "Improvements in an HMM-based Speech Synthesizer", *Proceedings of EUROSPEECH'95*, Spain, pp.573-576, 1995.

[5] "Whistler: A trainable Text-to-Speech System", *International Conference on Spoken Language Processing*, 1996.

[6] Sanghun Kim and J.C.Lee, "Korean Text-to-Speech System Using TD-PSOLA," *in Proc. SST94*, pp.587-592, 1994.

[7] J.C.Lee, and Sanghun Kim, and Minsoo Hahn, "Intonation Processing for Korean TTS Conversion Using Stylization Method," in Proc. *ICSPAT95*, pp.1943-1946, 1995.

[8] Sanghun. Kim, Hangsup Lee and Hoi R. Kim, "An Effectiveness of Automatic Labeling using Speech Recognizer", *SICOPS96*, SESSON 3.6, 1996.

[9] 성철재, 김상훈,"경계(Boundary) 신호의 지각적/음성적 분석-운율구 단위설정과 관련하여", 한글학회, 한글 232, 1996.

[10] Eric Sanders and Paul Taylor, "Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis", *Proceedings of EUROSPEECH'95*, Spain, pp.1811-1814, 1995.

[11] 김상훈, 성철재, 이정철,"운율구 경계현상 분석 및 텍스트에서 운율구 추출", 한국음향학회 제 16 권, 제 1 호, pp24-32, 1997.

[12] Sanghun Kim, Jung-chul Lee, and Jun Park, "Standardization of Korean ToBI system and Autolabeling Using Low-High Intonation Stylization", Proceedings of ICSP'97, Seoul, Korea, pp. 161-165, 1997.