운율 및 길이 정보를 이용한 무제한 음성 합성기의 설계 및 구현

양 진 석† · 김 재 범† · 이 정 현††

요 약

Text-to-Speech 시스템에서 자연스럽게 음성을 합성하기 위해서는 운율과 길이에 대한 처리가 선행되어야 한다. 이를 위해서, 자연어 처리에 의해 분석된 문장들에 대해 억양 규칙을 적용한 후, 반복적인 실험을 통해 운율 및 길이 정보를 추출하였다. 본 논문에서는 이러한 정보를 이용하여 Text-to-Speech 시스템에서 자연성을 향상시킬 수 있는 방법을 제안한다. 실험 결과, 본 논문에서 제안하고 구현한 무제한 Text-to-Speech 시스템이 이러한 정보들을 사용하지 않는 시스템과 비교해서 더 자연스럽게 문장들을 합성해 낸다는 것을 보였다.

Design and Implementation of a Text-to-Speech System using the Prosody and Duration Information

Jinseog Yang† · Jaebeom Kim† · Junghyun Lee††

ABSTRACT

To produce more natural speech in a Text-to-Speech system, the processing of the prosody and duration must be preceded. For this, we applied a sequence of intonation rules to the sentences analyzed by natural language processing in advance, and then extracted the prosody and duration information by means of trial-and-error experiments. In this paper, a method is proposed to improve the naturalness in a Text-to-Speech system using this information. As the results, the Text-to-Speech system proposed and implemented in this paper showed more natural speech synthesis than the systems, which do not use this information, did.

1. 서 론

음성 합성 시스템을 개발하는 데 있어서 중요하게 고려해야 할 사항은 명료성(intelligibility)과 자연성(naturalness)이다. 특히, 한국어를 위한 음성 합성 시스템에 있어서 명료성은 자음이나 모음에 따라서 단어의 의미가 바뀔 수 있고, 자연성은 컴퓨터로 하여금 보다 인간다운 발성을 가능케 한다는 점에서 각각 중요한 역할을 한다[1,2,6]. 그러나 명료성과 자연성은 서로 상반되는 경향을 보이고 있다. 즉, 명료성이 향상되면 자연성이 떨어지고, 자연성이 향상되면 명료성이 떨어지게 된다. 그러므로 이러한 두 성질을 최대한 절충해서 합성하는 방법을 사용하는 것이 바람직하다. 두 성질을 절충하기 위한 방법으로는 여러가지

※본 연구는 인하대학교 95년도 연구비 지원에 의하여 수행되었음.

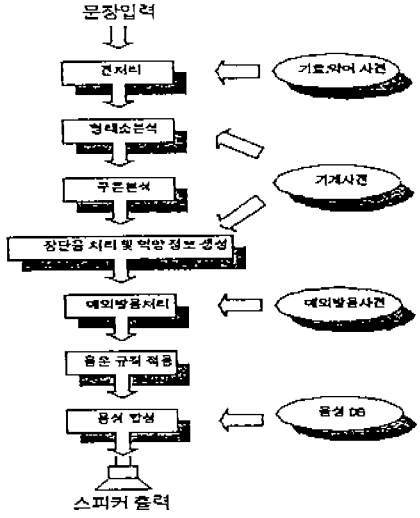
† 준 회 원:인하대학교 전자계산공학과 석사과정
†† 종신회원:인하대학교 전자계산공학과 교수
논문접수:1995년 12월 21일, 심사완료:1996년 5월 16일

가 있겠지만 여기에서는 '초성 +중성'(CV)형과 '중성 +종성'(VC)형의 음절 단위로 음성 DB를 결합하는 방법을 사용하여 명료성을 유지하고, 명료성이 유지되는 상황하에서 자연성을 향상시키는데 중점을 두고 있다.

자연성에 영향을 미치는 요소에는 여러가지가 있지만[2], 그 중에서 가장 큰 요소로는 운율과 모음의 길이 변화를 들 수 있다. 그러므로 본 논문에서는 일련의 억양 규칙에 의하여 운율 및 길이 정보를 추출하고 이를 실제 시스템에 적용하여 자연성을 향상시키는 방법을 제시한다.

2. 음성 합성기의 구성

한국어의 합성을 위한 Text-to-Speech 시스템은 (그림 1)의 전체 구성도와 같이 크게 7단계로 구성된다.



(그림 1) Text-to-Speech 시스템의 전체 구성도
(Fig. 1) Block diagram of Text-to-Speech System

2.1 전처리

전처리는 입력된 문장에 있어서 한글이 아닌 숫자, 약어, 기호 등을 모두 한글로 바꿔주는 역할을 한다. 이러한 처리는 한글로 입력된 문장만 합성이 가능하도록 제한을 두었기 때문이다. 그리고 숫자인 경우는 숫자뒤에 수량이나 단위를 나타내는 명사가 오는 경우와 오지 않는 경우에 따라 읽는 방식이 다를 뿐만 아니라, 수량이나 단위를 나타내는 명사에 따라 서로 읽는 방식이 달라지므로 이를 고려해서 적절하게 변환을 해야 한다. 예를 들어, 숫자 '5'에 대해 생각해 보면, 단위가 오지 않는 '5'는 '오'로 읽으며, '5번'은 '오번'으로, '5개'는 '다섯개'로 읽는다.

(그림 2)에서는 문장을 합성된 음성으로 출력하기 위해 COMPUTER와 DOS, 그리고 3이 전처리되어야 한다.

나는 COMPUTER에 DOS를 3번이나 설치했다.
→ 나는 컴퓨터에 도쓰를 세번이나 설치했다.

(그림 2) 전처리
(Fig. 2) Preprocessing

2.2 형태소 분석

형태소란 의미를 가진 최소의 단위를 말한다. 예를 들어, '사람이'라는 어절은 '사람'이라는 형태소(명사)와 '이'라는 형태소(주격 조사)로 구성되어 있다. 모든 문장은 이러한 형태소들로 분리될 수가 있는데, 이처럼 문장에서 형태소들을 분리하는 과정을 형태소 분석이라 한다. 본 논문의 실험에서는 크게 전편집 단계, 가능한 결합 추정 단계, 1차 필터링과 불규칙 처리 단계, 2차 필터링과 후편집 단계의 4단계로 구성된 형태소 분석기를 사용하였다[9].

(그림 3)의 예에서, '가장'이라는 단어는 '여럿 가운

[입력 문장]
멀티미디어는 최근 정보산업 분야에서 가장 많이 거론되는 분야이다.

[형태소 분석 후]
[멀티미디어 ((멀티미디어 는) (N PS))]
[최근 ((최근) (AD))]
[정보산업 ((정보산업) (N))]
[분야에서 ((분야 에서) (N PCA))]
[가장 ((가장) (AD) (가장) (N))]
[많이 ((많이) (AD))]
[거론되는 ((거론 되 는) (NH AS ED))]
[분야이다. ((분야 이 다) (N SFP EE))
((분야 이 다) (N IDA EE))]

(그림 3) 형태소 분석
(Fig. 3) Morphological analysis

데 어느 것보다 더'라는 의미를 가지는 부사 성분(AD)과 '집안의 어른', '남편', 혹은 '거짓으로 꾸밈'을 나타내는 명사 성분(N)으로 분석될 수 있다.

2.3 구문 분석

형태소 분석을 하게 되면 하나의 단어에 대해서도 여러개의 후보 형태소 결과들이 나타나게 된다. 그러므로 형태소 분석 결과에서 나온 여러 후보 형태소들 가운데 문법과 의미에 맞는 하나의 형태소를 선택해 주는 구문 분석이 필요하게 된다. 하지만 아직 구문 분석이 실험단계이므로 본 실험에서는 형태소들의 가능한 품사의 결합을 일련의 규칙으로 만들어 후보 형태소의 개수를 줄인 후, 나머지 후보 형태소들 가운데 결합 가능성이 가장 높은 하나를 선택하는 방법을 사용하였다. 즉 (그림 3)의 예에서 보면 '가장'이라는 단어는 부사성분(AD)과 명사성분(N)이 있는데, 이중 명사 성분은 뒤에 '많이'라는 부사성분(AD)이 올 수 없으므로 제거된다. 또한 '분야이다'에 대한 후보 형태소도 두개가 존재하는데, 이 경우는 둘다 앞의 관형형 전성어미(ED) 뒤에 명사가 나올 수 있으므로 두 후보 형태소가 적법하고, 이때는 임의로 하나의 후보를 선택하게 된다. (그림 4)는 (그림 3)의 형태소 분석에 대한 구문분석 결과를 보여준다.

```
멀티미디어(N)는(PS) 최근(AD) 정보산업(N) 분야(N)에서(PCA) 가장(AD)
많이(AD) 거론(NH)되(AS)는(ED) 분야(N)이(IDA)다(EE).
```

(그림 4) 구문 분석
(Fig. 4) Syntactic analysis

2.4 장단음 처리 및 억양 정보 생성

적절히 선택된 형태소들 중 명사인 경우는 이미 장단음 정보를 포함하고 있는 기계사전을 탐색하여 장음에 대해서는 명사 뒤에 장음 정보를 추가한다. 그리고 품사 정보에 따라 음운 단어, 음운구, 억양구등으로 나누어 적절한 표시를 하게 되는데, 음운 단어는 하나의 음운구를 구성하는 단어들로서 단어와 조사, 어미로써 구성되며 경계가 되는 음절(경계음)까지 붙여 읽어야 하는 단어들을 말한다. 음운구는 주부, 술부, 부사구등과 같은 통사적 정보를 반영하며 휴지가 삽입되는 단위이다. 또한 억양구(기식구)는 사람이 한번 숨을 들여마신 후 발음할 수 있는 하나 또는 그 이상의 어절들을 말하며 어떤 문장에 대해 전체의 억양 패턴을 결정하는 것으로 대개 문장의 끝이 되며, 이는 주로 문장 부호(., !, ?)에 의해 결정된다.

(그림 5)는 이러한 사항들을 고려하여 앞의 예문을 처리한 결과를 보여주는 것으로, 여기에서 : 표시는 장음 정보를 나타내며, +는 음운 단어, /는 음운구, 그리고 #은 억양구를 각각 나타낸다. 즉, '멀티미디어', '정보산업', '거론'은 모두 명사로 기계사전을 탐색한 결과 '멀:', '산:', '거:'에서 각각 장음으로 등록이 되어 있으므로 장음 표시가 추가되었지만, 명사 '분야'는 그러한 정보가 없으므로 그대로 사용되었다. 그리고 두개의 어절로 구성된 '정보산업 분야에서', '가장 많이', 그리고 '거론되는 분야이다'는 다음과 같은 몇가지 규칙에 의해 한번에 발성하게 되는 음운구가 된다.

① 명사 뒤에 조사 또는 어미가 오는 경우는 +를 삽입한다.
② 부사 뒤에 부사, 형용사, 동사가 오는 경우는 +를 삽입한다.
③ 형용사 뒤에 명사가 오는 경우는 +를 삽입한다.
④ 관형형 어미 뒤에는 +를 삽입한다.
⑤ 위의 네가지 경우에 속하지 않고 띄어쓰기를 한 경우에는 /를 삽입한다.
⑥ 네 어절 이상이 한 음운구를 이루는 경우는 제일 뒤의 어절을 따로 분리한다.
⑦ 문장의 끝에는 #을 삽입한다.

```
멀:티미디어(N)+는(PS)/ 최근(AD)/ 정보산:업(N)+ 분야(N)+에서(PCA)/
가장(AD)+ 많이(AD)/ 거:론(NH)+되(AS)는(ED)+ 분야(N)+이(IDA)다(EE).#
```

(그림 5) 장단음 처리 및 억양 정보의 생성
(Fig. 5) Processing of long/short sounds and generation of intonation information

2.5 예외 발음 처리

음성 합성시 자연성을 높이기 위해서는 음절 단위의 문자표기를 직접 음성으로 합성하기 보다는 사람이 실제로 발음하는 표기로 변환해서 합성하는 것이 바람직하다. 그러나 현재 한국어의 음운 규칙으로는 (그림 6)과 같은 단어들에 대해서는 올바른 표기가 불

가능하므로 이를 해결하기 위해 예외 발음 사전을 두고 음운 규칙을 적용하기 전에 먼저 이들 단어들을 알맞은 표기로 바꾼다.

본 논문의 경우는 대부분 한자어와 한국어 발음 사전[11]에서 예외 발음으로 정한 단어들을 중심으로 약 1,000여개의 단어로 구성된 예외 발음 사전을 구축하여 사용하였다.

사건 ⇒ 사껀
특성 ⇒ 특썽
할증 ⇒ 할쯩

(그림 6) 예외 발음되는 단어들
(Fig. 6) Exceptionally pronunciated words

2.6 음운 규칙의 적용

예외 발음이 아닌 단어들은 한국어 표준 발음 규칙에 해당하므로 이 단어들에 대해서는 문교부에서 정한 규칙[11]에 의해 변환하여 표기한다. 그래서 앞의 예문은 (그림 7)에서와 같이 '정보산업'은 '정보사넙'으로, '분야'는 '부냐'로, 그리고 '많이'는 '마니'로 바뀌게 된다.

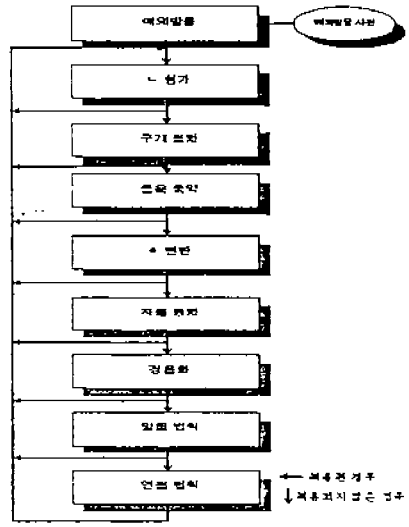
빌:티미디어+는/ 최근/ 정보사:넙+ 부냐+에서/ 가장+ 마니/
거:론+되는+ 부냐+이다.#

(그림 7) 음운 규칙의 적용
(Fig. 7) Application of phonological rules

적용되는 음운 규칙으로는 ㄴ 첨가, 구개음화, 음운 축약, ㅎ 변환, 자음동화, 경음화, 말음법칙 및 연음법칙이 있으며, 이들 규칙들이 적용되는 순서는 (그림 8)과 같다. 그리고 규칙 적용시에는 음운 규칙들을 여러 번 반복할 수도 있는데, 이는 여러가지 음운 규칙들이 복합적으로 적용될 수 있다는 것을 의미한다. 그래서 임의의 어절에 대해 음운 규칙이 하나라도 적용되면 그 어절은 변환되었으므로 처음부터 다시 음운 규칙을 적용하고, 이 과정을 반복해서 이후에 어떠한 음운 규칙도 적용되지 않으면 적용 과정을 끝낸다.

2.7 합성기

음성 합성을 하는 방법에는 조음(articulatory) 모델



(그림 8) 음운 규칙의 적용 순서도
(Fig. 8) Flowchart for applying phonological rules

방식, 포만트 합성(formant synthesis) 방식 등이 있다[5]. 조음 모델은 구현이 매우 복잡하고 계산량도 많은 단점이 있으며[1,2], 포만트 합성 모델은 자음을 합성해 내는데 있어 명료성이 떨어진다[2,4]는 단점이 있다. 그래서 본 논문의 실험에서는 '초성 +중성'(CV)형과 '중성 +종성'(VC)형의 음절 단위로 녹음된 음성 데이타를 결합하여 명료성을 유지하고, 이를 기반으로 하여 자연성을 향상시키기 위해 운율 및 길이 정보를 적용하는 합성기를 사용한다.

3. 운율 및 길이 정보의 추출

운율은 발화 단위에 내포된 의미정보에 따라 결정되는데 각 음절의 강세뿐만 아니라 구와 절의 위치, 기능, 상호결합관계 등의 복합적인 요인에 영향을 받게 된다[10]. 그러므로 이러한 요인들을 고려하여 운율 및 길이 제어 정보를 추출하는 것이 바람직하다.

(그림 9)는 입력 문장으로부터 운율 및 길이 정보를 추출하여 수치화한 결과이며, 추출된 정보들은 〈길이 정보, 피치 변경 정보〉의 형태로 각 음절의 앞에서 합성기에게 정보를 제공한다. 여기서 주의할 것은 장음 표시만 제외하고 음운 단어 표시, 음운구 표시, 억양구 표시는 수치정보와 함께 그대로 남아있다는 것이

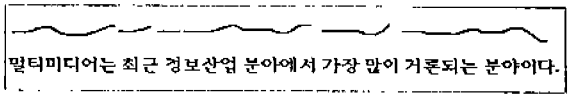
다. 그래서 음운 단어 표시(+)에 대해서는 중간에 어떠한 휴지 구간도 들어가지 않으며, 음운구 표시(/)와 억양구 표시(#)에 대해서만 휴지 구간이 들어가게 된다. 휴지 구간은 문장 중간의 음운구보다는 문장과 문장을 구분을 하는 억양구의 휴지 구간이 더 긴 것이 보다 자연스럽다. 본 논문에서는 실험을 한 결과 음운구에는 250msec정도의 휴지 구간을 두고, 억양구에서는 500msec정도의 휴지 구간을 두는 것이 자연스러운 것으로 나타났다.

<5,1>멀<2,1>티<3,0.99>미<3,0.99>디<3,1>어+<5,0.97>는/ <4,0.98>최<4,0.97>근/ <4,0.98>정<2,0.98>보<4,0.98>사<5,0.97>넘+ <3,0.98>부<4,0.98>나+<3,0.99>에<5,0.98>서/<3,0.99>가<4,0.99>장+ <3,1>마<5,0.98>니/ <4,0.99>거<4,0.99>론<3,1>되<4,0.99>는+ <3,1>부<4,0.98>나+<3,0.99>이<5,1.04>다.#

(그림 9) 운율 및 길이 정보를 수치화한 결과
(Fig. 9) Results of expressing the prosody and duration infomation numerically

운율을 제어하는데 있어 실제로 영향을 미치는 피치 변경값은 (그림 11)과 같은 규칙을 기반으로 하였으며, 규칙에서 사용된 수치는 반복된 실험에 의해 결정되었다. 여기서 P(0)는 이전 단위의 끝음절의 피치이며, P(1)은 첫음절을 나타내고 P(n)은 n번째 음절을 나타낸다.

이상과 같이 추출된 운율 및 길이 정보는 음성 합성기에서의 파형을 변환시켜 자연성을 높이게 되는데, 위의 정보에 따라 나타나게 되는 합성음의 억양 패턴을 나타내면 (그림 10)과 같다.



(그림 10) 억양 패턴
(Fig. 10) Intonation pattern

3.1 억양 규칙

운율 정보를 추출하기 위해서 [8]과 [10]으로부터 얻어진 결과를 종합하여 다음과 같은 억양규칙을 정의하였다.

①어떤 음절에 강세가 오면 그 다음 음절부터 피치가 올라간다.

②문두와 문중에는 억양구의 끝에서 피치가 최고조가 된다.

③문두와 문중에는 한 억양구가 끝난 후 임시 휴지기를 갖는다.

④문미 끝의 억양구는 그 문장의 유형에 따라 피치 유형이 결정된다.

⑤강세는 억양의 시작을 나타낸다.

⑥하나의 발화 단위 안에서 각 억양구의 최고 피치는 발화 끝으로 갈수록 낮아진다.

⑦억양구 안에 있는 음절의 수가 많을수록 각 음절이 차지하는 시간이 짧아진다.

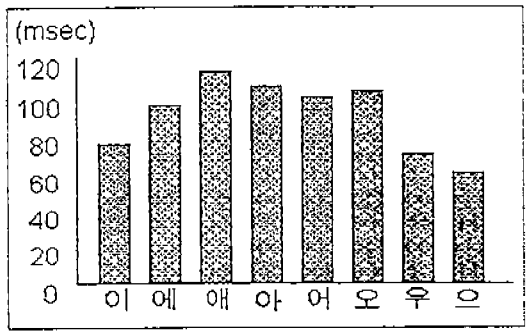
1. *1* 음절인 경우
   ① 문장의 시작 : P(1) = 1.00
   ② 음운 단어, 음운구 : P(1) = P(0) + 0.01
   ③ 억양구 : P(1) = P(0) + 0.02
2. *2* 음절 ~ (*n*-2) 음절
   ① P(*i*) = P(*i*-1) + 0.005 (2 ≤ *i* ≤ *j*)
   ② P(*i*) = P(*i*-1) - 0.005 (*j* < *i* ≤ *n*-2) (*j* : 강세 음절, 2 ≤ *j* ≤ *n*-2)
3. (*n*-1) 음절
   ① 음운 단어 : P(*n*-1) = P(*n*-2) - 0.005
   ② 음운구, 억양구 : P(*n*-1) = P(*n*-2) + 0.005
4. *n* 음절
   ① 음운 단어, 음운구, 억양구가 각각 문장의 첫번째일 경우
      P(*n*) = P(*n*-1) - δ
      (δ : 음운 단어 → 0.01, 음운구 → 0.03, 억양구 → 0.04)
   ② 그외의 경우
      음운 단어 : P(*n*) = P(*n*-1) - 0.005
      음운구 : P(*n*) = (이전의 최대 피치값 - 0.04)와 (P(*n*-1) - 0.005)를 비교하여 하향 흐름이 벗어나지 않는 것으로 선택.
5. 문장 부호가 .인 평서문인 경우 1.04로 설정한다.
6. 문장 부호가 ?인 의문문인 경우 0.95로 설정한다.

(그림 11) 피치 변경값 결정 규칙
(Fig. 11) Determining rules for the pitch values

3.2 모음에 따른 고유 지속 시간

모음의 길이는 조음 기관의 조음 시간으로 결정이 된다. 일반적으로 조음 기관의 이동거리가 짧은 경우(고모음 /이/가 해당)는 큰 경우(저모음 /아/가 해당)보다 길이가 짧으며, 움직임이 빠른 조음 기관(혀끝이 해당)으로 발음되는 경우가 움직임이 느린 조음기관(입술이 해당)으로 발음되는 경우보다 짧다. 또한 조음 기관의 긴장이 없는 연음(lax)의 경우는 긴장이 있는 경음(tense)보다 짧은데, 이는 경음을 발음하기 위해서는 조음 기관의 근육의 긴장이 이루어지는 시간이 소요되기 때문이다. 이처럼 모든 음성학적 조건

이 같을 때는 조음적인 제약에 의해 말소리의 길이가 달라질 수 있으며, 이것을 말소리의 고유 지속 시간(intrinsic duration)이라고 한다. 그러나 실제 말소리의 길이는 앞뒤의 소리, 발화의 길이, 악센트, 위치, 속도 등의 영향을 받아 나타나게 되는데, 이것의 상대적인 차이를 나타내면 (그림 12)와 같다[12].



(그림 12) 모음 지속 시간의 상대적 길이
(Fig. 12) Comparison of relative length among vowels

3.3 운율 및 길이 정보의 생성

형태소 분석과 구문 분석 후에 부가된 음운 단어, 음운구, 억양구의 표시에 따라 위의 억양 규칙을 적용하여 음성 사전의 피치를 변경할 정보를 수치로 표시한다. 수치값은 반복된 실험에 의해 본 논문의 시스템에 맞도록 정해진 것으로, 피치 변경값의 수치는 1.0을 기준으로 하여 피치가 낮은음을 만들 때는 1.0보다 큰 값을, 그리고 피치가 높은음을 만들 때는 1.0보다 작은 값을 사용한다. 그리고 길이 정보는 장단음 처리를 통해 얻은 명사의 장단 정보와 억양 규칙에 따른 길이 변화량을 고려하여 1에서 5까지의 정수값으로 표시하되 가장 긴 길이 정보를 5로 하였으며, 1에서 5의 각 단계에 대해 모음 길이에 있어 20msec 정도의 차이를 두었다.

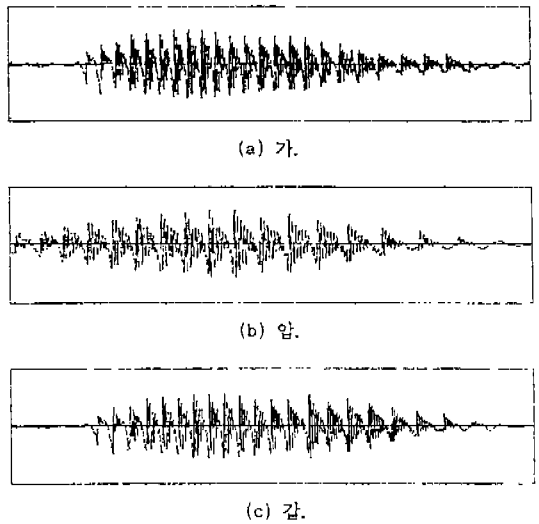
4. 음성 파형의 제어

4.1 음성 사전의 구축

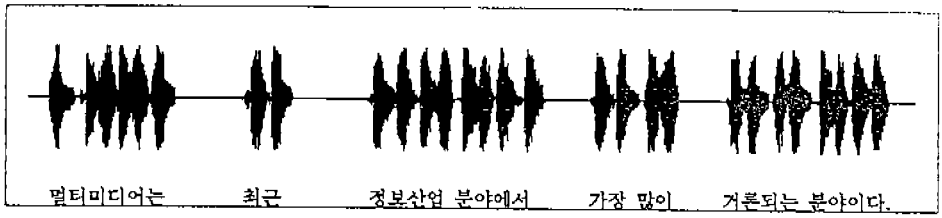
음성 사전은 한국 전자 통신 연구소(ETRI)에서 학술용으로 공개한 음성 DB를 사용하였다. 이 음성 DB는 KBS 아나운서에 의해 발성된 여성 음성 사전과 전문 발성가에 의해 발성된 남성 음성 사전의 두가지로 된 것으로 16-비트/16kHz로 연속 녹음되었으나, 실험에서 사용하기 위해 음절별로 hand-labeling을 하고 PC에서 사용이 가능하도록 화일의 저장 포맷을 변환하였다. 또한 1,244개로 구성된 데이타들 중에서 2음절로 된 데이타들을 뺀 466개의 단음절 데이타만을 선택하여 사용하였다. 그러나, 모든 음절을 합성하기 위해서는 초성 18, 중성 22, 종성 7개가 있으므로 CV형 396개 +VC형 154개 = 550개의 단음절 데이타가 필요하므로 DB에 누락된 음절에 대해서는 파형 편집기를 이용하여 기존의 파형을 조합하여 사용하였다.

4.2 파형의 결합

실험을 위해 구축된 음성 DB는 두개의 음소로 결합된 글자들만을 사용하였다. 즉, '초성 자음 +중성 모음'으로 구성된 글자들과 '중성 모음 +종성 자음'으로 구성된 글자들을 발성한 파형들로 구성된 DB이다. 이 DB로부터 '초성 +중성 +종성'으로 구성된 글자를 만들기 위한 방법으로는 주파수 영역에서 음소에 대한 파라미터를 추출하여 결합하는 방식[1,2]과 시간 영역에서 두개의 파형을 결합하는 방식[2]이 있는데, 본 논문에서는 시간 영역에서 두개의 파형을 결



(그림 13) (a)(b) 원래의 음성 DB. (c) 결합된 형태.
(Fig. 13) (a)(b) Original speech DB. (c) Concatenated waveform



(그림 14) 한 문장에 대해 결합된 파형의 결과
(Fig. 14) Result of concatenating waveforms for a sentence

합하는 방법을 사용하였다. 그리고 파형을 결합하는 위치는 모음 부분으로 하였는데, 이는 파형에서 모음이 차지하는 부분이 대부분이고 거의 주기성(quasi-periodic)을 가지기 때문이다.

예를 들어, (그림 13)은 두개의 음소로 구성된 '가'와 '압'을 사용하여 세개의 음소로 구성된 '갑'의 파형을 보여주며, (그림 14)는 한 문장 전체에 대해 결합된 파형을 보여준다.

4.3 운율 및 길이 정보에 따른 파형의 변환

추출해 낸 정보를 이용한 운율 제어에는 Multirate Interpolation/Decimation 필터를 사용하였으며[6], 길이의 제어에는 수치 정보와 모음의 고유 지속 시간에 따라 모음 중간의 피치를 삽입 또는 제거하는 방법[7]을 사용하였다. 그리고 앞에서 문제가 된 세개의 음소로 구성된 글자에 대해서는 먼저 '초성+중성'으로 구성된 단어에 대해 피치 변경 정보 만큼 필터링을 한 후에 '중성+종성'으로 구성된 글자에 대해 필터링을 하였으며[8], 그 다음에 두글자를 결합하는 방법을 사용하였다.

피치를 변경하는 수치는 그 범위를 0.90과 1.10사이로 한정하였다. 이 범위는 실험에 의해 정해진 값으로 이 범위를 벗어나면 스펙트럼의 왜곡이 심해져서 음색이 변하게 된다. 특히 여성의 경우는 이러한 현상이 두드러지게 나타나므로 이 범위의 최대/최소값인 0.90과 1.10도 음색을 조금 변화시키는 결과를 보였다.

5. 실험 및 평가

본 청취실험에는 남자 10명과 여자 5명의 서울·경기 지방의 표준말을 사용하는 합성음에 경험이 없는 피험자들이 참여하였다. 제시된 문장은 국민교육헌장을 공통으로 합성하여 들려주었고, 피험자들이 신문 사설에서 직접 선택한 10개의 문장을 운율 및 길이 정보를 사용하여 합성한 음성과 그러한 정보를 사용하지 않고 단순히 음운 규칙만을 적용하여 합성한 음성을 들려준 후 평가하였다.

5.1 실험 평가의 방법

- 실험환경: 한국어 처리 및 음성 합성을 위해 실험에서 사용한 시스템으로는 75MHz의 CPU가 장착된 펜티엄, D/A 컨버터로는 Sound blaster-16, 합성용 음성 DB로는 남성 화자의 발성음을 사용하였다.
- 평가용 청취자군의 선정: 합성음에 경험이 없고 서울말을 사용하는 남자 10명, 여자 5명으로 선정하였다.
- 평가용 자료: 우선 국민교육헌장을 공통으로 선정했고 신문 사설에서 피험자가 무작위로 선택한 문장으로 10개씩 선정하였다. 이때 피험자가 문장을 읽고 선택하게 되면 청취 시 예측이 가능하므로 명료도 평가에 영향을 미치게 된다. 따라서 '몇일자 신문의 몇번째 문장'과 같은 방법으로 선택하게 하였다.
- 평가 실험: 자연음을 10점으로 하여, 운율 및 길이 제어를 하지 않은 합성음과 이들을 제어한 합성음을 0에서 10사이의 점수를 주도록 하여 평균한 MOS(Mean Opinion Score) 평가를 하였다.

5.2 실험 결과

평가 실험의 결과는 〈표 1〉과 같다. 실험 결과 명료

성은 9.7점 정도의 높은 점수를 받음으로써, 피험자군이 합성음의 의미를 정확히 파악한 것으로 나타났고, 자연성은 운율 및 길이 제어를 하지 않은 합성음에 비해 0.9점 정도 향상되었다고 평가되었다.

〈표 1〉 실험 결과
〈Table 1〉 Results

	운율 및 길이 제어를 하지 않은 합성음		운율 및 길이 제어를 한 합성음	
	자연성	명료성	자연성	명료성
남자(10명 평균)	6.1	9.8	6.8	9.8
여자(5명 평균)	5.5	9.7	6.6	9.6
평 균	5.8	9.75	6.7	9.7

5.3 각 모듈별 오류도

형태소 분석 모듈의 오분석은 'ㄹ' 탈락시 발생하고 있으며 470개의 문장을 분석한 결과 0.5%의 오류도를 보인다. 구문 및 의미 분석 모듈은 후보 형태소가 하나 이상 남았을 때 그중 확률적으로 높은 하나의 형태소를 선택하는 데서 발생하며 이는 약 11%의 오류도를 보인다.

6. 결론 및 향후 과제

실험 결과 '초성 +중성'형과 '중성 +종성'형의 음절단위로만 문장들을 합성했기 때문에 명료성은 매우 만족할 수준이었으며, 자연성 향상에도 많은 효과가 있었다. 또한 기존의 합성 방법들 중에서 포만트 합성 방식은 자연성이 뛰어난 반면에 명료성이 떨어지며, 하나의 완전한 음절 단위로 모든 문장을 합성해내는 합성법은 명료성은 최대가 될 수 있으나 음성 DB가 커진다는 단점을 고려해 볼 때, 본 논문에서 제안한 방법은 이러한 문제들을 적절히 해결한 방법이라고 할 수 있다. 또한 한국어의 운율특성을 일련의 규칙으로 만들고 계속된 반복 실험을 통해 수치화 함으로써 운율제어에 의한 자연성 향상에도 많은 효과가 있었다.

향후 과제로 한국어 운율특성을 보다 정확히 분석하여, 현재는 다소 단순화된 운율규칙을 좀더 세분화해야 하며, 음소의 지속시간에 대한 규칙도 아울러 세분화되어야 할 것이다. 이에 의해 보다 정확한 운율 및 길이 정보가 추출되고, 음절간의 연결 부분에 대한 변음구간 처리가 적절히 이루어진다면 이러한 Text-to-Speech 시스템들에서의 자연성은 상당히 향상될 것이라고 기대된다.

참 고 문 헌

[1] Dennis H. Klatt, "Software for a cascade/parallel formant synthesizer," J.Acoust. Soc. Am, Vol. 67, No. 3, pp. 971-995, 1980.

[2] Dennis H. Klatt, "Review of text-to-speech conversion for English," J.Acoust. Soc. Am, Vol. 82, No. 3, pp. 737-793, 1987.

[3] J.H.McClellan, T.W.Parks, L.R.Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," IEEE Trans. Audio and Electroaucoust., Vol. AU-21, No. 6, pp. 506-525, Dec. 1973.

[4] J.L.Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, 1972.

[5] L.R.Rabiner, Bernard Gold, *Theory and Application of Digital Signal Processing*, Prentice Hall, 1975.

[6] L.R.Rabiner, R.W.Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[7] M.J.Ross, H.L. Shaffer, A.Cohen, R.Freudberg, H.J.Manley, "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. ASSP. Vol. 22, pp. 353-362, Oct. 1974.

[8] 구희산, "음성 합성의 운율처리를 위한 악센트 연구," 음성·음운·형태론연구, 한국문화사, pp. 21-34, 1993.

[9] 여상화, *다단계 필터링을 이용한 형태소 분석기의 설계 및 구현*, 인하대학교 공학석사 학위 논문, 1992.

[10] 임홍빈, "국어 억양의 기본 성격과 특징," 새국어생활, 제3권 1호, pp. 58-89, 1993.

[11] 전영우, *표준 한국어 발음사전*, 집문당, 1992.

[12] 지민제, 최운천, 김상훈, "우리말 소리의 길이:실험 음성학적 연구," 제5회 한글 및 한국어 정보처

리 학술발표 논문집, pp. 119-130, 1993.

양 진 석

1995년 인하대학교 전자계산공학과 (학사)
1995년~현재 인하대학교 전자계산공학과 석사과정
관심분야: 음성신호처리, 자연어처리

김 재 범

1995년 인하대학교 전자계산공학과 (학사)
1995년~현재 인하대학교 전자계산공학과 석사과정
관심분야: 음성신호처리, 자연어처리

이 정 현

1977년 인하대학교 전자공학과 (학사)
1980년 인하대학교 전자공학과 (공학석사)
1988년 인하대학교 전자공학과 (공학박사)
1979년~1981년 한국전자기술연구소 시스템 연구원
1984년~1988년 경기대학교 조교수
1989년~현재 인하대학교 전자계산공학과 교수
관심분야: 자연어처리, 음성신호처리, HCI