고품질 음성합성을 위한 합성 DB 구축

강동규, 이승훈, 류원호
주식회사 코아보이스

Speech Database Design and Structuring for High Quality TTS

Dong-Gyu Kang, Sionghun Yi, Won-Ho Ryu
CoreVoice, Inc.
E-mail : dgkang@corevoice.com

Abstract

As the telematics service that is the integration of information technology approaches commercialization, the necessity and gravity of speech technology is rapidly growing. The speech technology occupies important position in the telematics service because it informs the starting of service and the retrieved result. This service must provide high accuracy of speech recognition and natural synthesis of human speech in a driving environment and it is especially true for the fee-for-service.

For high quality TTS, the speech synthesis technique that makes optimal synthesis database and uses efficiently this database is required. In this paper, we describe the design of phonetically balanced sentences used for speech database, the selection of service-suitable-speaker, the extraction methods of accurate phoneme boundary, and the factors which are taken into consideration in the extraction stage of prosody. Finally we show the real case that has commercially implemented.

I. 서론

음성기술이 미래의 10대 기술로 선정되는 등 산업사회에서 유망기술로 대두되고 있지만 아직까지 음성기술은 대부분 부가 서비스나 전체 서비스 시스템에서 없어도 되는 보조기능을 담당해 왔다. 최근 수년전부터 준비해 왔던 서비스 시스템 구축을 완료하고 본격적인 서비스를 시작하고 있는 텔레매틱스는 자동차 내에서 생활하는 시간이 증가하면서, 자동차 내에서 보다 편리한 서비스를 받고자하는 욕구에서 시작된 서비스이다. 대도시의 복잡한 도심지에서 원하는 목적지까지 혼잡한 구간을 피하여 최단시간에 차량으로 이동할 수 있는 지능형 위치 안내 서비스, 차량고장 진단과 동시에 가장 가까운 서비스 센터까지 안내해 주는 서비스, e-mail과 각종 공공정보 안내 서비스 등을 비롯한 갖가지 새로운 서비스들이 준비될 전망이다. 이 서비스에서 음성기술은 서비스의 시작과 처리된 결과를 최종적으로 고객에게 전달하는 없어서는 안될 중요한 역할을 담당하고 있다. 또한 대부분이 유료 서비스를 시행하고 있으며 그 가입자 수는 높은 증가 추세를 나타내고 있다.

텔레매틱스에 필요한 음성기술은 자동차 환경에서의 음성인식과 고품질의 음성합성기술이 요구되고 있다. 특히 음성합성의 경우 자동차 잡음과 운전에 집중한 상태에서도 정확하고도 편안하게 알아 들을 수 있을 정도의 자연성과 명료도가 요구된다.

고품질의 합성음을 얻기 위한 방법으로서 원래의 음성데이터를 가능한 그대로 사용하려는 방법들이 많이 이용되어 왔다. 초기에는 합성단위에 대한 후보를 1개로 구축하고 PSOLA방법으로 운율을 변경하여 사

용하였으나 억양과 성도길이에 보상이 어려워 고품질을 얻지 못하였다. 이에 대한 대안으로 합성단위 후보를 복수 후보로 구축하고 일부 운율을 변경함으로써 고품질을 실현하였다.[1]

운율을 변경하는 것은 통계적 운율정보를 바탕으로 하고 있으므로 자연스런 원래 화자의 운율을 반영하기는 매우 어려운 방법이다. 원래 화자의 자연스런 운율을 반영하기 위해서는 풍부한 운율정보를 포함할 수 있을 정도의 합성 데이터베이스를 구축하고, 원음을 가공하지 않으면서 정밀한 자연운율을 예측할 수 있어야 한다.

본 논문에서는 코퍼스 기반 음성합성 시스템에서 고품질 합성음을 얻기 위한 합성 DB 문장 (Phonetically Balanced Sentence) 설계, 화자선정, 음소경계 추출, 억양정보 추출에 있어서 고려해야할 사항들에 대해 분석하였고, 실제로 구축한 상용화 사례를 제시하였다.

II. 합성 DB 문장 설계

고품질의 합성 DB 구축을 위한 문장은 먼저 풍부한 음운환경에 의한 모든 음소가 포함되어야 하고 각 음소별로는 다양한 운율이 담긴 음성데이터가 필요하다. 이를 위해서는 대용량의 말뭉치 데이터를 발음변환하여 음운환경의 분포를 고려한 문장을 추출해야한다.[2]

문장 DB는 크게 나누어 일반영역에서 추출한 문장과 서비스 영역별로 추출된 문장으로 구분된다. 서비스 영역에는 일기예보, 교통정보, 증권정보, 위치정보, 관광정보, 의료정보, 법률정보, 첨단 과학 기술정보, 부동산정보 등이 있으며 이는 특정영역의 서비스에서 보다 자연스런 합성음을 얻기 위한 것이다. 일반영역의 경우 다양한 음운환경을 고려하여 추출하며 서비스 영역별 문장은 다양한 운율에 초점을 맞추어 추출하는 것이 바람직하다.

하나의 합성단위는 비교적 많은 후보로 구성되는 것이 바람직하므로 청취에 의해 구분이 어려운 유사음의 경우에는 통합하여 보다 많은 후보를 확보하는 것이 합성음의 안정성을 유지할 수 있다.

III. 화자선정

구축된 PBS는 단순한 문장일 뿐 실제적으로 음성합성에 필요한 것은 문장을 낭독한 음성데이터이다. 문장으로부터 음성데이터를 얻기 위해
서는 비교적 발음 훈련이 잘되어 있고, 서비스에 적합한 음색을 가진 전문 성우나 아나운서의 도움으로 얻을 수 있다.

먼저 개략적인 발음과 서비스에 적합한 선호도를 갖는 화자를 선택하기 위해 수십 명의 후보 화자가 낭독한 음성데이터를 이용하여 다수의 청취자가 후보를 선정한다. 선정 시 고려사항은 운율의 안정감, 발음의 정확성, 음성의 명료도, 음색의 선호도 등이다.

운율의 안정감은 화자가 유사한 문장을 다시 읽었을 경우 유사한 운율을 낼 수 있는 화자가 안정된 운율특성 재현에 매우 유리하며, 발성속도가 일정해야 합성음이 안정될 수 있다.

발음의 정확성은 몇 가지 문장으로 판단하기 어렵기 때문에 최소한 난이도 있는 문장으로 수백 문장을 녹음해야 보다 정확한 판단이 가능하다.

음성의 명료도는 화자가 발성 시 활기 있고 힘차게 발음할 수 있어야 많은 문장을 녹음할 경우 음색이 변하지 않고 명료한 음성을 얻을 수 있다. 또한 전화망에서 사용할 경우 전화선로에 의한 손실로 인하여 명료도가 낮아지지 않는 음성이 적합하다.

음색의 경우, 서비스 용도에 따라 약간씩 다를 수 있으며 일반적으로 상냥한 말씨를 선호하는 경향이 있다. 그러나 이 경우에는 운율의 변화폭이 크므로 합성음의 안정성이 저하될 수 있다.

IV. 음소경계 추출

정확한 음소분할은 합성단위 간 스펙트럼의 연속성이 보장되고 정확한 지속시간 및 강세 정보를 모델링하거나 예측이 가능하므로 합성 DB 구축에서 가장 중요한 부분 중에 하나이다.

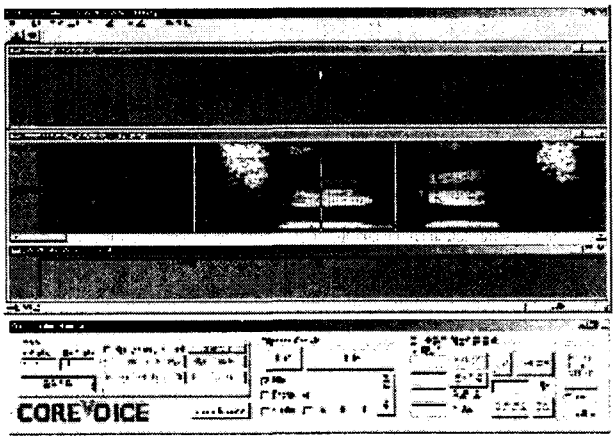

그림 1. 음소분할 도구의 예

일반적으로 코퍼스 기반의 합성 DB를 구축하는 경우에는 대용량이므로, 작업시간을 줄이기 위하여 훈련된 여러 사람이 동시에 음소분할을 수행한다. 이러한 경우에는 음소분할 경계에 대한 기준이 서로 다르므로 일관성을 유지하기 어려워 합성 품질이 저하된다.

이를 극복하기 위해서는 음소분할을 수행하는 사람들이 일정한 기준을 가질 수 있도록 주기적인 훈련과 교육이 필수적으로 병행되어야 한다.

또한 음소분할 도구 역시 일관성을 유지할 수 있도록 설계되어야하고, 음성청취에 의해 음소경계를 판단할 경우 다양한 방법으로 청취할 수 있어야 보다 정확한 경계를 추출할 수 있다.

V. 억양정보 추출

코퍼스 기반의 합성음 품질은 다양한 음운환경에 대한 합성단위의 종류와 각 후보들에 대한 풍부한 운율후보가 갖추어질 경우 높은 품질의 합성음을 얻을 수 있다. 그러나 이러한 풍부한 운율 후보가 존재하더라도 이에 대한 정확한 특징 추출이 선행되지 않는다면 원래화자의 자연운율을 재현할 수 없으므로 합성음의 자연성과 안정성은 크게 저하되게 된다.[3]

운율특성은 그 중요도에 따라 억양, 지속시간, 강세특성으로 나타나며 지속시간과 강세특성은 정밀 음소분할시 정보를 추출할 수 있으나 억양정보의 경우 별도의 처리가 필요하다.[4] 억양정보를 보다 편리하게 추출하기 위해서는 문장 녹음 시 EGG신호를 동시에 녹음하면 보다 정확하면서도 편리하게 억양정보를 추출할 수 있다. 이 경우 화자의 목 부분에 센서를 부착해야 하므로 화자가 발성시 약간의 불편이 수반되어 발성에 정확도가 저하 될 수 있다. 또한 화자가 이를 거부하는 경우도 있으므로 자세한 설명이 필요할 경우도 있다.

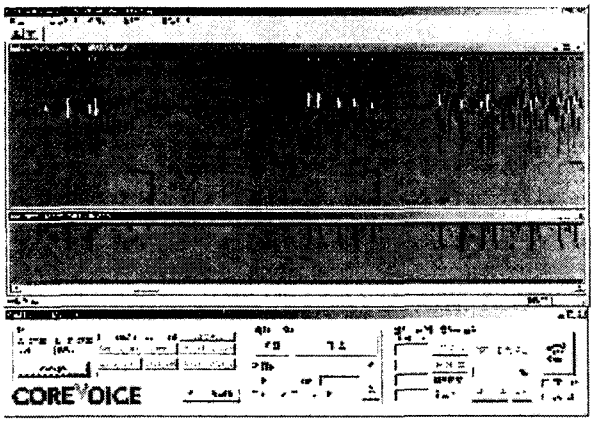

그림 2. 억양정보 추출 도구의 예

EGG를 이용하더라도 실제 운율과 추출된 운율정보와 일치하지 않는 경우가 있다. 즉 EGG에 의해 추출된 억양정보의 정확도는 90~95%정도이며 나머지 5~10%의 오류는 훈련된 사람에 의해 오류를 수정해야한다

대부분의 운율에 있어서 억양정보는 매우 중요한 정보로서 약간의 오류가 발생하여도 부자연스럽게 느껴지므로 반드시 검증을 해야 할 부분이다.

VI. 구현 및 결과

6.1 최적 합성 DB 구축을 위한 문장 설계

다양한 장르로 구성된 약 100만 문장의 말뭉치에서 음운환경과 음소의 분포를 고려하여 일반영역 문장을 약 5000문장 추출하였다. 문장을 추출하는 방법으로는 대량의 코퍼스에 나타난 음운 환경과 음소 분포를 효율적으로 나타낼 수 있는 방법을 사용하였다.[5]

이와는 별도로 다양한 분야에서 합성음의 품질을 최대화하기 위해 일기 예보에 사용된 멘트, 교통 상황 안내 멘트, 증권 정보, 부동산 정보, 관광지 안내 정보, 법률 상담 정보, 위치 안내 정보, 약품 및 의료 정보, 첨단 과학 기술 관련 정보, 그리고 각종 ARS에 사용된 멘트들을 수집하여 5000문장을 추출하였다.

이들 문장에서 추출된 고유한 합성단위 수는 55000개 정도에 이르고 최대 복수 후보 수는 약 11000개에 이른다.

각 합성단위에서 유사 음운환경을 통합하여 약 45000개로 정도로 축소하였다. 합성단위의 종류는 성우의 발음 정확도에 따라 그 숫자가 상당수 줄어 들 수 있다.[6]

6.2 화자선정

후보화자 50여 명 중에서 2~3개정도의 음성샘플을 청취시험에 의해 20~30대 남녀 10명이 운율의 안정감, 발음의 정확성, 음성의 명료도, 음색의 선호도에 대해 3회를 채점한 결과를 평균하여 높은 점수별로 3명을 선정하였다. 선정된 3명에 대해 구축된 문장 중에서 음운환경의 밀도가 높은 600문장을 각각 녹음하고 수동으로 음소분할을 수행하여 소규모 TTS 시스템을 제작하였다. 제작된 3명에 대한 소규모 TTS를 이용하여 다양한 문장을 합성한 다음 20~30대 남녀 10명이 합성음질을 평가한 다음, 최종적으로 1명을 선정하였다. 선정된 1명에 대하여 나머지 9400문장을 녹음하였다.

6.3 음소분할 및 운율정보 추출

먼저, 녹음된 10000문장에 대해 문장별로 음성을 분할하고 전사된 발음의 오류를 수정하였다. 그리고 음성인식기를 이용하여 자동으로 음소분할을 한 다음, 4명의 전문 레이블러 들이 약 4개월에 걸쳐 1회의 음소분할을 완료하고, 약 3개월에 걸쳐 억양정보 오류를 수정하였다. 이 결과로 구축된 음성합성 DB는 16kHz /16bit로 구성할 경우 약 3.0 GByte가 되었다.[7]

다수의 레이블러가 음소분할을 할 경우에는 음소분할 기준에 대한 일관성 유지가 매우 어렵기 때문에 일주일에 2~3회 정도의 주기로 지속적인 음소분할 기준에 대한 교육을 병행하여 일관성을 최대한 유지할 수 있도록 하였다. 억양정보의 오류 유형은 원래 억양 보다 2배수가 되거나 반수가 되는 경우, 음성 에너지가 작아서 억양을 느끼지 못하지만 급격한 변화가 나타나는 경우 등이 있었다.

6.3 음질 평가

3.0 GByte급으로 구성된 TTS 시스템의 합성음의 평가를 위해 20~30대 남녀 20명에 의해 MOS 테스트를 수행하였다. 시험 문장은 합성DB 구축 시 사용된 문장과 연관성이 전혀 없는 문장과 연관성이 있는 문장으로 구성하였다.

합성DB문장과 연관성이 있는 문장의 경우 자연음과 구별하기 어려운 정도의 음질로서 MOS 4.0 정도의 수준이었고 연관성이 없는 경우에는 다소 음질이 저하되어 MOS 3.5 정도의 등급을 얻을 수 있었다.

VII. 결론

본 논문에서는 코퍼스 기반 음성합성 시스템에서 고품질의 합성음을 얻기 위한 문장설계, 화자선정, 음소 경계 및 억양정보 추출에 있어서 고려해야 될 사항들에 대해 분석하였고 실제 구축한 상용화 사례에 대하여 기술하였다.

코퍼스 기반 합성시스템의 장점은 특정영역에서는 고품질을 얻을 수 있으나 다양한 영역에서는 품질이 저하되는 것과 DB의 크기가 매우 크다는 단점이 있다. 영역에 관계없이 고품질을 얻기 위해서 문장설계, 화자선정, 음소 및 억양정보의 정밀한 추출 등을 고려하여 초대형 합성 DB를 구축함으로써 상용화에 충분한 고품질의 합성음을 얻을 수 있었다.

컴퓨터의 처리속도 개선과 메모리의 집적도 향상에 힘입어 대용량 DB(1.0~2.0 GByte)를 이용한 상용화에는 많은 문제가 해소되고 있지만 초대용량 DB(2.0 GByte 이상)를 이용한 상용화에는 아직도 어려움이 남아 있다. 또한 초대용량의 경우 음소경계 및 운율정보 추출에 많은 인력과 시간이 소요될 뿐만 아니라 여러 사람에 의한 일관성 저하로 여전히 품질이 낮아지는 문제점들이 남아 있다. 이를 극복하기 위해서는 음소경계 및 억양정보 추출과정을 보다 정확하게 처리할 수 있는 자동화방법이 개발되어야 할 것이다.

참고문헌

[1] Nick Campbell, Alan W.Black, *Progress in Speech Synthesis*, pp.279-292, springer, 1996

[2] Steve Young and Gerrit Bloothooft, *Corpus-based methods in Language and Speech Processing*, Kluwer Academic Publishers, 1997

[3] P. Price et al., "The use of prosody in syntatic disambiguatin," J.Acoust. Soc. Amer., vol. 90, pp. 2956-2970, 1991

[4] Allen J., Hunnicutt S., and Klatt D. *From text to speech: the MITalk system*, MIT Press, Cambridge, Massachusetts, 1987

[5] Hsin-min Wang, *Statistical Analysis of Mandarin Acoustic Units and Automatic Extraction of Phonetically Rich Sentences Based Upon a Very Large Chinese Text Corpus*, Computational Linguistics and Chinese Language Processing, vol. 3 no. 2, pp.93-114, August 1998.

[6] Black, A., and Taylor, P., "Automatically clustering similar units for unit selection in speech synthesis," *Proc. Eurospeech*, pp.601~604, Sep 1997

[7] Donovan, R.E., *Trainable Speech Synthesis*, PhD. Thesis, Cambridge University Engineering Department, 1996