LSP를 이용한 음소단위 PSOLA 음성합성에 관한 연구

A Study on Phoneme-Based PSOLA Speech Synthesis Using LSP

권 혁 제*, 조 순 계**, 김 종 교*
(Hyuck Je Kwon*, Soon Kye Cho**, Chong Kyo Kim*)

요 약

본 논문에서는 음소단위 PSOLA 한국어 합성을 LSP line의 조절과 자모음 분석을 통해서 실시하였다. 음성합성에서 많이 사용하는 triphone, diphone, demisyllable등과 같은 합성단위들은 자연스러운 합성음을 위해 다양한 음운환경에서 수집된다. 그러나, 이런 방법은 많은 시간과 메모리가 요구된다. 본 논문에서는 합성단위로서 자음17개, 모음16개로 총 33개의 음소를 이용하였다. 자음은 후위모음/이/인 CV에서 segment되고, 모음은 단음절의 단모음과 이중모음을 1인의 화자로부터 합성데이터를 수집하였다. 또한, 10명의 화자가 발성한 CV에서 각 모음에 따라 변하는 자음의 주파수를 분석하였고, CV +VC 또는 CV +CV에서 각 자음에 따라 변하는 모음의 포먼트변화를 분석하였다. 분석결과를 토대로 모음은 LSP line을 조절해서 PSOLA합성을 하고, 자음은 합성하려는 모음과 결합하였다. 그 결과 총 6개의 합성단어에 대한 청취율은 65%를 보였다.

ABSTRACT

A PSOLA synthesis of Korean speech based on phoneme units is performed by a modification of LSP line and an analysis of consonants and vowels. For the synthesized speech to be natural, the synthesis units such as triphone, diphone, demisyllable, etc. should be collected in various prosodic environments. It requires too much of time and memory.

In this study, we used 33 phoneme units composed of 17 consonants and 16 vowels as the synthesis units. The consonant phonemes are obtained from the CV(C +이) type words and the vowel phoneme from single vowels and diphthongs, uttered by a speaker.

For the CV type words uttered by 10 speakers, the frequency energy of consonants versus the subsequent vowels is analyzed. For the CV +VC and CV +CV type words, the formant frequency variation of the vowels versus the preceding or subsequent consonants is also analyzed. The vowels' PSOLA synthesis is performed by a modification of LSP line and the consonants are combined with the vowels to be sysnthesized. The experimental results shows the recognition accuracy of 65% for 6 synthesized words.

Ⅰ. 서 론

합성기에서 주로 사용되는 기본 합성단위는 diphone, demisyllable, triphone과 syllable단위등이 있다. 이런 단위들은 천이구간의 조음정보를 그대로 유지한 상태에서 각기 가지고 있는 환경에 맞게 조합을 하면 된다. 하지만, 여러 음운환경에 따른 데이터를 수집해야하며, 메모리가 필요하게 된다. 이런 메모리단점을 극복하고자 합성단위를 음소단위로 하여 구현하였다. 음소단위 합성기의 구현은 음소가 가지고 있는 성질 이외에 각 음운환경에 대한 적응성도 고려해야 한다. 현재 음소단위 합성기는 그 청취력이 낮은 편이지만 메모리문제와 수집의 문제를 해결하기 위해서는 필요하다고 하겠다. [1]

본 논문은 Ⅱ장에서는 데이터 수집 관련, Ⅲ장에서 각 자음에 따른 모음의 포먼트성분의 영향과 천이구간을 결정짓는 자음에 대하여 기술하였고, Ⅳ장은 음성합성방법인 PSOLA, Ⅴ장은 천이구간처리에 대하여 기술하였으며, 마지막으로 실험 및 평가를 Ⅵ장에서, 결론을 Ⅶ장에 기술하였다.

Ⅱ. 데이터

자음과 모음에 대한 분석은 화자 10명(20대 남자 10명, 전라북도)의 음성을 수집하여, 미리 정해진 발성목록을

*전북대학교 전자공학과
**조선대학교 공업전문대학교 전자통신학과
접수일자 : 1997년 5월 17일

보고 발성하게 하였다. 발성목록은 CV형태로 각 모음별 자음들로 구성하였다. 자음은 특정 알고리즘을 사용하지 않고 발성음에서 세그먼트해서 분석하였다. 모음의 경우는 포먼트의 변화를 알기 위해서 CV+VC 또는 CV+CV 형태의 발성형태로 발성하여 분석하였다. 합성에 사용할 자음은 후위모음에 대한 영향을 가능한 적게 하기 위해서 F1과 F2의 거리가 가장 먼 모음 /이/를 이용하였다. F1이 다른 모음의 것보다 가장 낮은 위치에 있고, 다른 모음에 비해 F2는 F1에 비해 그 에너지가 적기 때문에 합성시 /이/모음의 영향을 줄이면서 합성모음의 포먼트를 살렸다. 모음은 단모음과 이중모음을 수집하였다.

데이터를 구축하기 위하여 컴퓨터의 fan 소리와 키보드소리, 작은 소리의 대화등이 있는 실험실 환경하에서 Dynamic mike를 사용하여 음성처리보드인 speech station으로 음성을 수집하였으며, 대상 화자로는 20대중반 남자1명의 발성을 사용하였다. 이때 데이터의 샘플링은 10kHz, 16bit, integer linear PCM방식을 사용하여 저장하였으며, 음성처리보드의 offset을 제거하기 위하여 일정한 크기의 integer 데이터로 산술연산을 하였다. 화자는 학습을 통해 훈련된 음성을 발성하게 하였으며, 훈련 내용은 정확한 주파수 대역이 나타나도록 하였다. 이렇게 수집된 음성데이터는 local peak wave가 손상이 가지 않도록 임의 크기의 Hamming창을 씌워서 구축하였다.

Ⅲ. 자음분석

일반적인 특징으로 모든 자음은 후위모음에 따라서 주파수가 나타나는 위치가 변한다. 즉, 모음의 포먼트성분 중 F2에 따라 변한다. [2] 표 1은 각 모음에 대한 자음의 주파수 변화를 나타내었다. 자음분석구간은 모음의 주파수에너지를 포함하지 않는 구간을 세그먼트해서 분석해 FFT한후에 각 대역별 에너지를 구하고, 그 중 가장 큰 값으로 다른 대역의 값을 나눈 비율이다.

(1)모음의 포먼트성분들을 따라 현저하게 주파수의 에너지가 나타나는 자음들이다. 즉, 각 포먼 트성분의 상단부분에 현저하게 나타나며, /ㄱ, ㄷ, ㅂ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ/등이 있다. /ㅊ, ㅋ, ㅌ, ㅍ/와 같은 격음은 시간영역에서 power가 크게 나타나며, 저주파에서 에너지가 집중 되어 나타나기도 한다.

(2)/ㅅ, ㅈ, ㅊ/의 경우는 마찰음으로 3kHz이상에서 에너지가 집중되는 현상이 보인다. 또한 모음의 주파수 변화에 따른 F2에서의 변화도 나타난다.

(3)모음에 따라 변화하는 다른 자음들과는 달리 /ㅁ/은 주파수가 나타나는 부분이 대략적으로 일정하다. 하지만 대부분 1kHz이하의 같은 주파수 대역에 /ㅁ/의 주파수가 나타난다.

표 1. 모음에 대한 자음의 주파수 특성(주파수에너지/max. 주파수에너지)
Tabel 1. The frequency rate of consonants for vowels (energy/max. energy) in frequency

	ㄱ							ㄷ						
kHz	아	어	오	우	으	이	에	아	어	오	우	으	이	에
0-0.5	0.876	0.9047	0.9739	1	0.8681	0.8576	0.8937	0.95	1	1	0.9621	1	1	1
0.5-1	0.908	0.9273	1	0.9715	0.9215	0.7705	0.8082	0.914	0.9951	0.9932	0.9432	0.8733	0.7968	0.8075
1-1.5	0.982	1	0.8182	0.8289	1	0.7589	0.817	0.932	0.9953	0.9952	0.9574	0.9039	0.8751	0.7388
1.5-2.0	0.963	0.9347	0.7554	0.771	0.8784	0.792	0.8606	0.899	0.9939	0.9945	0.9057	0.8992	0.869	0.717
2.0-2.5	0.87	0.847	0.7317	0.7629	0.8258	0.8361	0.9078	0.857	0.9948	0.9945	0.9426	0.8976	0.8476	0.7929
2.5-3.0	0.861	0.8935	0.7363	0.8278	0.8203	0.936	0.9905	1	0.9955	0.9951	0.9898	0.9359	0.874	0.7805
3.0-3.5	0.94	0.8704	0.7657	0.8134	0.8456	1	1	0.903	0.9969	0.995	1	0.9349	0.9684	0.8466
3.5-4.0	0.929	0.9124	0.8463	0.952	0.832	0.9434	0.9457	0.935	0.9933	0.9943	0.8979	0.9353	0.9025	0.653
4.0-4.5	1	0.9726	0.8673	0.9461	0.898	0.8935	0.9551	0.945	0.9949	0.9925	0.9455	0.9398	0.9327	0.6639
4.5-5.0	0.767	0.7405	0.6078	0.682	0.7183	0.7056	0.7259	0.729	0.8397	0.8414	0.6987	0.7104	0.7328	0.5283

	ㅂ							ㅂ						
kHz	아	어	오	우	으	이	에	아	어	오	우	으	이	에
0-0.5	0.946	1	1	1	1	1	1	1	1	1	1	1	1	1
0.5-1	0.817	0.7426	0.7512	0.7528	0.959	0.7841	0.7427	0.939	0.9356	0.9119	0.8657	0.8611	0.8916	0.863
1-1.5	0.884	0.7671	0.8507	0.8061	0.9568	0.7763	0.764	0.907	0.9112	0.8252	0.8098	0.7869	0.8209	0.8162
1.5-2.0	0.858	0.704	0.7098	0.7158	0.9508	0.7591	0.7602	0.852	0.8112	0.8352	0.7575	0.7691	0.8869	0.8555
2.0-2.5	1	0.725	0.7277	0.7965	0.9788	0.8646	0.8304	0.84	0.8689	0.8335	0.7464	0.7676	0.8233	0.8067
2.5-3.0	0.965	0.7363	0.7424	0.7513	0.9692	0.9151	0.8613	0.932	0.9025	0.8379	0.7761	0.7646	0.94	0.863
3.0-3.5	0.952	0.8052	0.8463	0.8479	0.9642	0.7821	0.717	0.852	0.8821	0.8431	0.8141	0.8009	0.8595	0.8034
3.5-4.0	0.977	0.6904	0.7266	0.7339	0.9406	0.6969	0.6859	0.809	0.8355	0.803	0.8621	0.8386	0.8387	0.8146
4.0-4.5	0.938	0.6998	0.7494	0.7992	0.9407	0.7103	0.6718	0.85	0.8805	0.8362	0.8371	0.8311	0.8198	0.7655
4.5-5.0	0.719	0.5681	0.591	0.585	0.7885	0.5666	0.5501	0.636	0.6703	0.6706	0.6065	0.6095	0.6139	0.6137

	ㅅ							ㅈ						
kHz	아	어	오	우	으	이	에	아	어	오	우	으	이	에
0-0.5	0.763	0.725	0.7564	0.7557	0.7919	0.6777	0.9553	0.788	0.8358	0.9475	0.8859	0.9445	0.7668	0.9042
0.5-1	0.735	0.6818	0.72	0.7316	0.7115	0.6345	0.9386	0.751	0.7629	0.8326	0.799	0.9337	0.721	0.8074
1-1.5	0.83	0.7335	0.7709	0.7863	0.8224	0.6692	0.9549	0.751	0.7734	0.7996	0.8034	0.9336	0.7015	0.7594
1.5-2.0	0.769	0.7512	0.7691	0.7796	0.7943	0.6433	0.9446	0.821	0.8218	0.8469	0.876	0.9433	0.7376	0.8057
2.0-2.5	0.885	0.7511	0.8101	0.8691	0.8122	0.7252	0.9736	0.836	0.8498	0.9802	1	0.9578	0.8519	0.8626
2.5-3.0	0.93	0.7679	0.8603	1	0.8945	0.8381	0.9864	0.807	0.8109	0.8513	0.9362	0.9507	0.7572	0.8103
3.0-3.5	0.925	0.831	1	0.9557	0.9541	0.873	0.9078	0.906	0.9232	1	0.9983	0.9659	0.8921	0.8878
3.5-4.0	0.931	0.8978	0.9796	0.8464	0.9497	0.9911	0.9859	0.956	1	0.9624	0.9716	0.9974	0.9869	0.94
4.0-4.5	1	1	0.901	0.7933	1	1	1	1	0.9988	0.9182	0.982	1	1	1
4.5-5.0	0.762	0.752	0.6665	0.5892	0.7696	0.7246	0.8128	0.774	0.7318	0.7118	0.7412	0.8186	0.7539	0.7262

	ㅊ							ㅋ						
kHz	아	어	오	우	으	이	에	아	어	오	우	으	이	에
0-0.5	0.996	0.8846	0.9975	1	0.944	0.8795	0.9849	0.959	1	1	0.9278	0.9847	0.913	0.9675
0.5-1	0.984	0.9	1	0.8171	0.9644	0.7934	0.9627	0.93	0.9344	0.9519	0.8572	0.895	0.8106	0.838
1-1.5	0.995	1	0.9785	0.8612	1	0.73	0.9464	0.929	0.9435	0.902	0.8397	0.87	0.7734	0.8643
1.5-2.0	1	0.8541	0.9666	0.8005	0.9679	0.8272	0.9598	0.96	0.9442	0.9023	0.8806	0.9176	0.7869	0.8954
2.0-2.5	0.983	0.8154	0.9677	0.8006	0.9592	0.8823	0.981	0.925	0.9259	0.9072	0.8806	0.9128	0.8626	1
2.5-3.0	0.982	0.8277	0.9615	0.8007	0.9561	1	1	0.971	0.9676	0.9275	0.9565	0.9418	0.9139	0.9724
3.0-3.5	0.993	0.8297	0.9658	0.8042	0.9554	0.9655	0.9878	1	0.9651	0.9869	1	1	0.9748	0.9877
3.5-4.0	0.992	0.8895	0.9667	0.8169	0.9581	0.9248	0.9773	0.952	0.9593	0.9566	0.8735	0.9863	1	0.9631
4.0-4.5	0.997	0.8792	0.9768	0.8173	0.9619	0.8957	0.9767	0.946	0.9739	0.8897	0.8447	0.9761	0.9786	0.9407
4.5-5.0	0.821	0.6591	0.8111	0.7055	0.8007	0.6853	0.8142	0.711	0.7728	0.6934	0.6843	0.7311	0.6972	0.7455

	ㅌ							ㅍ						
kHz	아	어	오	우	으	이	에	아	어	오	우	으	이	에
0-0.5	0.994	1	1	1	0.9843	0.9616	1	1	1	1	1	1	0.9721	1
0.5-1	0.987	0.9355	0.9022	0.8175	0.8654	0.8654	0.9656	0.897	0.9025	0.9539	0.8818	0.8964	0.8385	0.8449
1-1.5	0.985	0.9369	0.8537	0.8502	0.9173	0.8437	0.965	0.881	0.8319	0.95	0.8269	0.8569	0.8384	0.8074
1.5-2.0	0.998	0.8907	0.8818	0.8434	0.9281	0.9351	0.9788	0.83	0.8393	0.9443	0.8277	0.8488	0.9572	0.9029
2.0-2.5	0.986	0.8988	0.9032	0.8266	0.9317	0.8597	0.9768	0.878	0.8305	0.9452	0.8561	0.8526	1	0.9139
2.5-3.0	0.953	0.923	0.879	0.8585	0.9564	0.9879	0.9781	0.894	0.7928	0.9321	0.8613	0.8526	0.9919	0.9206
3.0-3.5	0.925	0.9293	0.9326	0.8627	0.9712	0.9237	0.9818	0.821	0.8147	0.9434	0.9025	0.8607	0.9881	0.8906
3.5-4.0	1	0.9639	0.9566	0.8714	1	1	0.9907	0.886	0.7701	0.9476	0.9321	0.8834	0.9138	0.8656
4.0-4.5	0.912	0.8901	0.921	0.8275	0.9338	0.9348	0.9786	0.816	0.7772	0.9395	0.8952	0.8486	0.9454	0.8246
4.5-5.0	0.695	0.6487	0.6627	0.7253	0.7358	0.6772	0.8116	0.63	0.5843	0.7628	0.6399	0.6362	0.65	0.6192

	ㅎ						
kHz	아	어	오	우	으	이	에
0-0.5	0.822	1	1	1	1	0.9419	0.9932
0.5-1	0.936	0.9281	0.89	0.8834	0.8012	0.7657	0.9606
1-1.5	0.926	0.878	0.8231	0.846	0.8584	0.7495	0.9528
1.5-2.0	0.881	0.8234	0.7842	0.7995	0.8488	0.8008	0.9754
2.0-2.5	0.932	0.8368	0.8211	0.7983	0.8032	0.8329	0.9906
2.5-3.0	1	0.893	0.7946	0.8131	0.8352	1	1
3.0-3.5	0.938	0.8742	0.7967	0.8265	0.8002	0.958	0.9827
3.5-4.0	0.99	0.9393	0.79	0.8761	0.8113	0.9659	0.9882
4.0-4.5	0.945	0.8376	0.7611	0.8087	0.8304	0.9558	0.9735
4.5-5.0	0.705	0.6669	0.5812	0.6044	0.6336	0.6855	0.7975

④대부분 유음에 많이 발생하는 경우로 F2의 상단부분에서 약 40Hz정도 윗 주파수 대역에서 모음의 F2로 하강곡선을 그리며 내려온다. 이런 모습은 /이, 외/와 단모음을 제외한 모든 단모음에 나타나는 두드러진 특징이다. /이/와 /외/는 특별히 F3부분이 /ㄴ/과 같이 발음할 경우에 주파수에너지가 증가한다. 이런 유음들로는 /ㄴ, ㄹ/이 있다.

⑤천이구간처리를 위한 기본적인 정보로 어느 자음이 모음에 영향을 주는냐 하는 것이다. 실험적인 방법으로 발성음의 스펙트로그램 분석과정를 거쳐 두드러지게 천이가 발생하는 자음을 수집한다. 이런 자음들로는 /ㄴ, ㄹ/과 /ㄷ/이 있다. /ㄷ/같은 경우에는 문장의 중간에만 발생하고 첫음으로 시작하는 곳에서는 천이가 발생하지 않는다. 반면에 /ㄴ, ㄹ/은 그렇지 않다.

Ⅳ. 합 성

4. 1 PSOLA 음성합성방식[7][8][10][11]

PSOLA방식은 시간영역에서 음성을 local peak에 따라 분해하여 그 파형을 원하는 피치의 패턴으로 재구성하는 합성 방법이다. 이 합성법은 주파수 영역과 달리 연산량이 적고, 피치의 조절이 간단하고 손쉽다는 특징이 있다. PSOLA방식은 음성을 분해한 후 다시 이 분해(분석)된 data를 가지고 재합성한다. 분석은 이미 추출한 pitch를 바탕으로 이루어진다. 사람이 음성을 인지할 때 중요한 역할을 하는 것은 각각의 local peak이다. 특별히 유성음이나 모음과 같은 경우에 연속적으로 반복되는 비슷한 형태의 파형이 발생되는데 이때 이 파형이 음성을 인지할 때 중요한 역할을 담당하게 된다. 이런 local peak가 손상되지 않게 하면서 음성을 분석해야 한다. Window를 사용해서 frame처리를 하게 되는데, 이때 window를 local peak를 중심으로 대칭적으로 분석하고, window의 크기는 각각의 pitch성분에 따라 조정이 된다. 즉, pitch 성분에 따라서 일정크기를 window에 적용한다. [10] 본 논문에서는 pitch의 두 배 만큼을 window 크기로 잡았다. 그림 4.1과 같이 window의 길이가 pitch에 따라 변화하는데, 이것은 pitch성분에 따라서 음성정보를 최대한 분석하기 위한 것이다.

· Analysis

$$S_m(n) = h_m(t_m - n)S(n) \tag{4.1}$$

$S_m(n)$: m번째 signal, $S(n)$: 원 음성파형

$h_m(n)$: Hamming window

t_m: m번째 pitch mark(or local peak)[8], [9]

· Hamming window

$$w(n) = \left(0.54 - 0.46\cos\frac{2\pi n}{N-1}\right) \quad N: \text{총 샘플수}$$

$$x_w(n) = w(n)x(n) \tag{4.2}$$

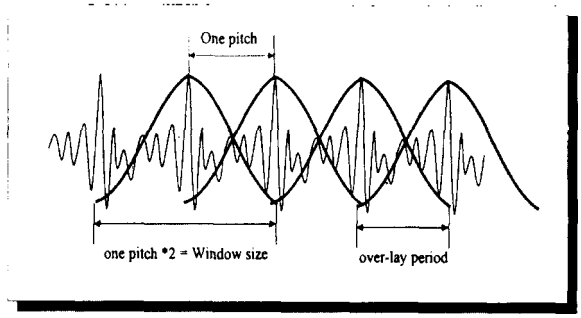


그림 4.1 PSOLA분석
Fig. 4.1 The analysis of PSOLA

4.2 재합성

분해를 통하여 각각 저장되어 있는 음편은 원음성의 local peak를 중심으로 분석하였기 때문에 재합성때에는 이 부분을 최대한 그대로 유지하면서 합성을 한다. 합성시에는 미리 정해져 있는 피치에 따라 합성이 이루어진다. 피치에 의해서 합성되는 것은 물론이고, 한 피치의 데이터 갯수에 따라서 합성하고자 하는 음의 장단도 결정된다. [9][12][13]

PSOLA합성방식의 특징이라고 할 수 있는 overlap-add는 연산량이 적고 구현하기 쉽다.

· Analysis

$$S_q(n) = S_f(n - t_q) \tag{4.3}$$

$S_q(n)$: q번째 합성파형, t_q: q번째 합성파형 pitch mark

· Over-Lap Add

$$S_{lap}(n) = S_f(N + n - O_s) + S_{f+1}(n)\ n = 0, 1, \cdots O_s \tag{4.4}$$

O_s: overlap 구간, N: 한 음편의 sample 수,

S_f: 분석 frame 파형 S_{lap}: overlap 구간의 합성파형

식(4.3)과 같이 각 음편은 서로 겹쳐지는 부분에 대해서는 단순한 더하기로 겹쳐진 구간을 처리한다. 또한 음이 이어지는 부분에서 에너지차이가 발생하게 되면 합성음은 낯설게 들리게 되기 때문에 에너지를 선형적으로 연결해야 한다. 앞음의 마지막 음편(프레임)과 뒤음의 첫음편의 에너지의 비를 구한 다음 뒷음의 첫 음편을 이 비로 조절한다. [8]

음운환경에 따라 음의 길이가 변한다. 길이가 변할 때 가장 두드러진 특징이라고 할 수 있는 것은 모음의 길이가 변한다는 것이다. 파찰음이나 파열음의 경우에는 그 변화가 심하게 나타난다. 그러므로 합성시 장단의 변화는 모음의 길이만을 조절한다. 모음을 도입부, 안정부, 종료부로 나누고 단음인 경우는 안정부를 삭제한다. [8][13] 각 환경에 따른 음의 길이는 정확한 자료가 부족해서 실

제 발음을 기준으로 그 길이를 정하였다. 또한 두 모음의 결합시 발생하는 피치의 차이는 각 음의 피치의 패턴은 변화시키지 않으면서 접합부분의 피치를 일치시켰다. 평서문의 경우는 첫음절의 피치가 가장 높고 뒤로 갈수록 낮아지는 경향을 보며, 의문문의 경우 뒤쪽이 상승하는 경향을 보인다.

V. 천이구간 처리

5.1 Line Spectrum Pairs(LSP)

Acoustical tube가 손실이 없고, 무한개의 공진점 Q를 가지고 있을 때, 스펙트럼의 공진을 나타내는 p차방정식의 근은 z 평면상의 단위원에 존재하는 $z_i = e^{j\lambda_i}$가 된다.

PARCO식은 다음과 같다.

$$A_{p-1}(z) = A_p(z) + k_p B_{p-1}(z)$$
$$B_p(z) = z^{-1}[B_{p-1}(z) - k_p A_{p-1}(z)] \quad (5.1)$$

모든 근들이 z-평면상의 단위원 상에 존재하기 때문에 식(5.1)을 각각 $p/2$ 차수인 $P(z)$, $Q(z)$로 분리하고, $P'(z) = \frac{P(z)}{(1+z)}$, $Q'(z) = \frac{Q(z)}{(1-z)}$ 를 단위원상에서 방정식을 계산한다. 여기서, $z = e^{j\omega}$이고 $z^1 + z^{-1} = 2\cos\omega$이다. [3][4][5]

$$P'(z) = e^{jp\omega/2}\left[A_0\cos\left(\frac{p}{2}\omega\right) + A_1\cos\left(\frac{p-2}{2}\omega\right) + \cdots + \frac{1}{2}A_{p/2}\right]$$
$$Q'(z) = e^{jp\omega/2}\left[B_0\cos\left(\frac{p}{2}\omega\right) + B_1\cos\left(\frac{p-2}{2}\omega\right) + \cdots + \frac{1}{2}B_{p/2}\right] \quad (5.2)$$

위의 식을 $x = \cos\omega$로 대치하면, LPC차수가 10차인 경우에는

$$P'(x) = 16A_0x^5 + 8A_1x^4 + (4A_2 - 20A_0)x^3 + (2A_3 - 8A_1)x^2 + (5A_0 - 3A_2 + A_4)x + (A_1 - A_3 + 0.5A_5)$$
$$Q'(x) = 16B_0x^5 + 8B_1x^4 + (4B_2 - 20B_0)x^3 + (2B_3 - 8B_1)x^2 + (5B_0 - 3B_2 + B_4)x + (B_1 - B_3 + 0.5B_5) \quad (5.3)$$

이다. [6] 식(5.3)의 두 방정식에 의해 얻어지는 10개의 근(pole점)들이 이루는 line을 조정함으로써 포먼트 성분을 조절한다. 그 주파수 성분인 line spectrum frequency는

$$LSF(i) = \frac{\cos^{-1}(x_i)}{2\pi T} \quad 1 \le i \le p \quad (5.4)$$

이다.

5.2 포먼트 조정

모음은 조음위치에 따라서 각기 나타나는 주파수 특성이 다르다. 모음 삼각도는 F1와 F2성분에 의해 분리하였으며, 혀의 조음위치에 대비하여 표현할 수 있다. 각 모음은 고유의 주파수 특성인 포먼트 성분을 가지면서 다른 모음으로 옮아가는 경우에는 그 주파수 특성이 변하면서 옮아가는 모음의 주파수 특성으로 변하게 된다. 이렇게 다른 모음으로 옮아가는 구간을 천이구간이라고 한다. 모든 현상이 다른 현상으로 옮아가게 되면 항상 거기에는 과도현상이 나타나듯이 이 천이구간도 이런 과도현상으로 음의 자연스러운 진행에 있어서 중요한 역할을 한다.

합성에서 천이구간처리는 합성음의 자연스러움을 보장하기 위한 필수적인 처리이다. 본 논문에서 사용하는 합성단위인 음소는 다른 합성단위 diphone이나 demisyllable, triphone들보다 음운환경에 대한 적응성이 떨어진다. 즉, 각 음운환경에 따른 음소를 선택하기 보다는 단음의 음소를 취해서 데이터를 구성하였기 때문에 환경이 바뀌게 되면 그에 상응하는 처리가 있어야만 한다. 모음은 발성음을 인지할 때 가장 중요한 역할을 담당하는 요소로서, 모음의 조절이 곧 논문에서 합성음의 질을 결정짓는 중요한 과정이다.

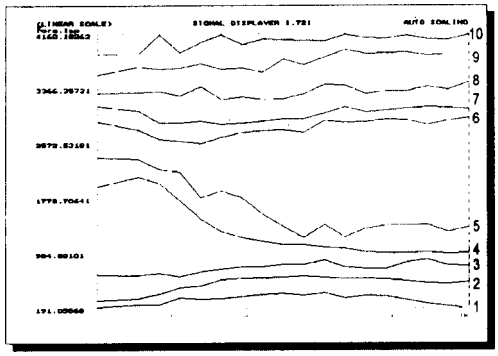
천이구간처리는 10차 LSP분석을 통해서 포먼트위치를 분석하고, 이를 근거로 LSP line을 조정한다. 두 LSP lines의 근접시리에 의해서 결정된 lines을 포먼트로 추정하고, 추정된 LSP lines을 LSF규칙에 어긋나지 않도록 조정한다. [14][15][16][17] LSP에 의해서 분석된 각 모음의 포먼트를 근거로 친이해야 할 길이와 높이(주파수)를 결정하게 된다. 이런 규칙은 실험적인 방법을 통하거나 분석을 통해서 이루어 진다. [3] 선행음들과의 관계는 앞서 기술한 것과 같이 대표음가로 발음되는 자음들의 영향에 의해서 선후모음의 일부 포먼트가 변한다. 즉, 대표음 /ㄷ/이 나타나는 경우, 대부분 모음은 F1과 F2사이가 벌어진다.

하지만, 대표음이 /ㄱ, ㅂ/인 경우에는 그렇지 않다. 조정이 끝난 전체 line은 다시 LPC로 변환후 원음성으로 바꾼다. 변환된 원음성을 PSOLA합성법을 이용해서 pitch와 음의 길이를 조절한다. 그림 5.1의 (a), (b)는 10차의 LSP분석으로 얻어진 /아/의 수정전후의 LSP line들이다. LSP 분석시 line의 수는 분석차수와 동일하며, 무음구간에서 각 line간의 간격은 샘플링 주파수와 연관되어 나타난다. 즉, 그림 5.1과 같이 샘플링 주파수가 10KHz인 경우 스펙트로그램상에는 5KHz가 나타나며, 10차의 LSP분석에 의해 각 line들은 등간격으로 배열되어 약 500Hz 정도의 차이를 보이면서 무음구간을 나타낸다.

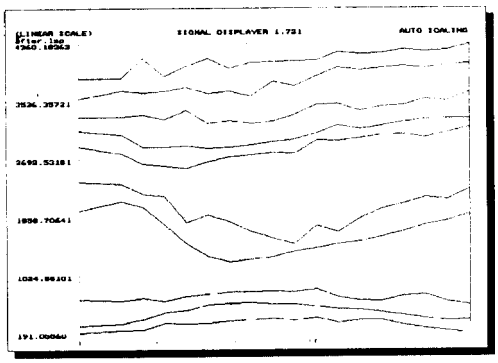
하지만, 유음구간에서는 각 line들이 서로 가까워지면서 주파수 에너지의 집중을 가져온다. 이렇게 집중되면서 포먼트성분이 구성되고, 음의 특성이 결정된다. 분석을 통해 나타난 line들은 실험을 통해서 pair를 결정짓게 되는데, 이때 pair를 이룬 두 line사이에 포먼트성분이 존재하게 된다. 그림 5.1에서 pair는 2와 3, 4와 5, 6과 7이다.

분석된 pair를 통하여 스펙트로그램상의 주파수변화를 조정할 수가 있다. 그림 5.1(b)는 그림 5.1(a)의 pair 성분

을 조정해서 얻어진 결과이며, 그림 5.2가 스펙트로그램으로 본 주파수 특성이다. 이중모음 /야/의 F2성분과 F1성분은 음의 앞부분에서 전이가 발생하다가 다시 합쳐졌으며, 이것을 LSP line조정을 통하여 뒷부분을 다시 천이시켰다.



(a) 수정전 LSP lines
(a) Source LSP lines



(b) 수정후 LSP lines
(b) Modified LSP lines

그림 5.1 수정전후의 LSP lines의 모습[3]
Fig. 5.1 Source and modified LSP lines

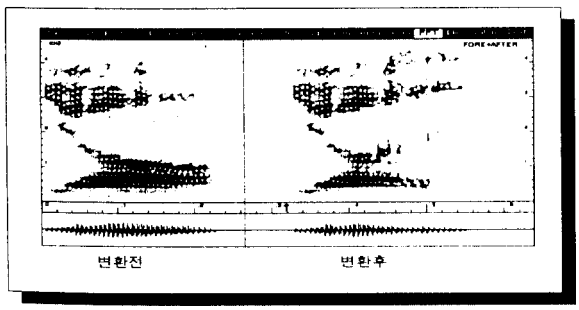
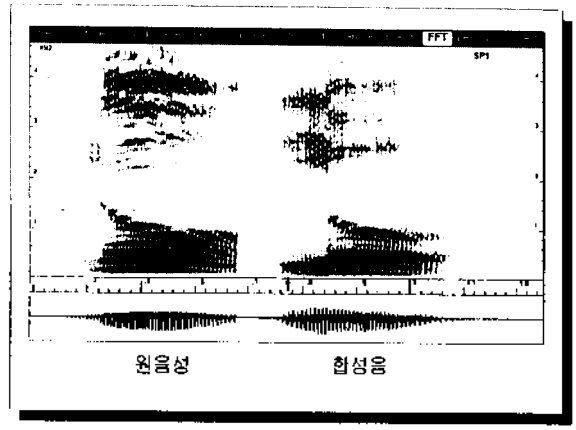


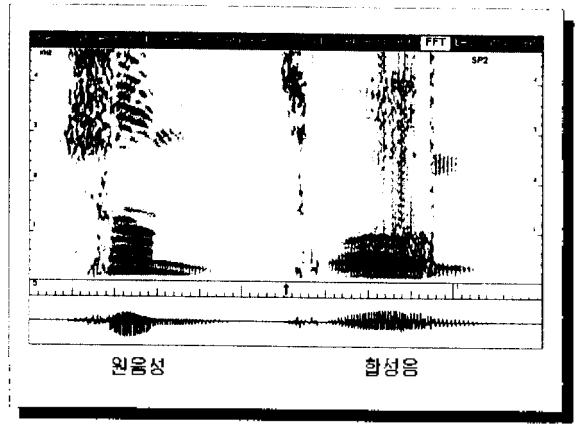
그림 5.2 LSP 수정전·후의 음성의 주파수 특성[3]
Fig. 5.2 The spectrogram of speech before and after modification

VI. 실험 및 평가

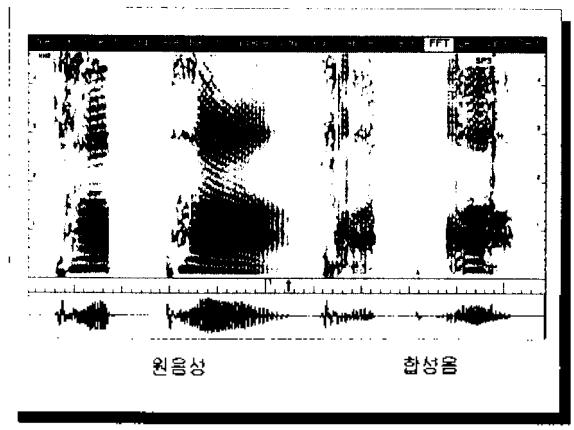
합성은 Pentium-PC 75MHz를 이용하여 처리하였고, 합성데이터를 DSP보드 TMS320C30으로 재생하였다. 합성음에 대한 평가는 합성단어에 대한 사전인지가 없는 대상자를 임의로 10명을 골라서 1음절, 2음절, 3음절 단어들을 들려주고 그 음에 대한 인지도를 측정하였으며, 이때의 청취율을 나타내면 표 1과 같다.



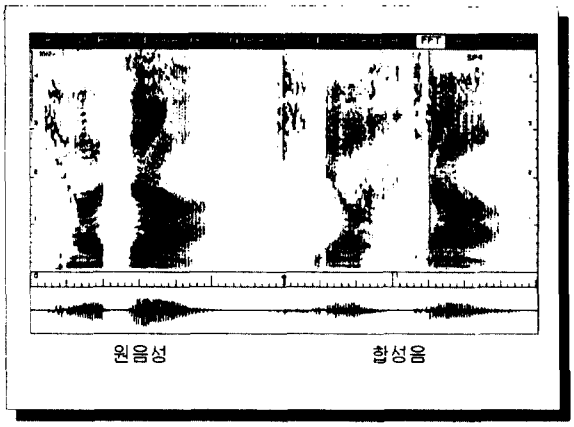
(a) /누/



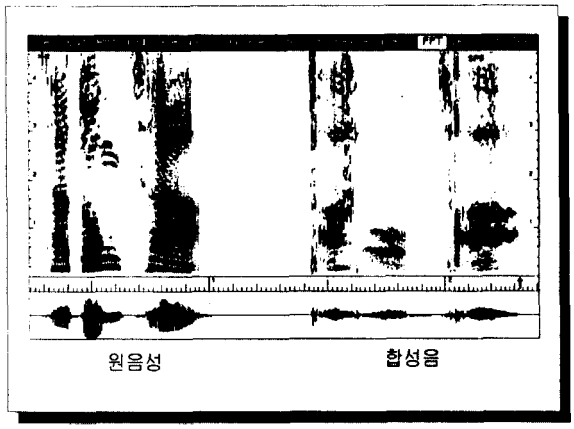
(b) /춤/



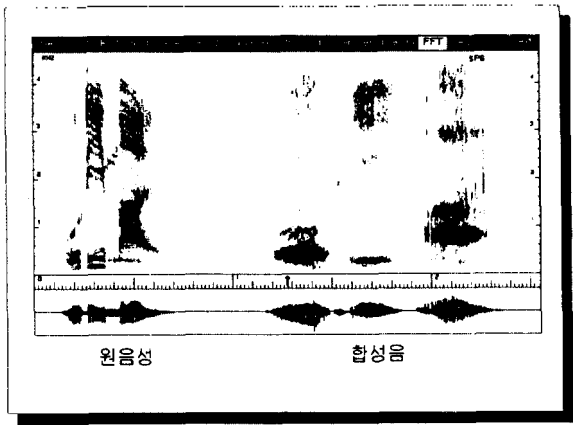
(c) /타파/



(d) /겨자/



(e) /자동차/



(f) /우리말/

그림 7.1 각 합성음절에 대한 원음성과 합성음의 스펙트로그램[2]
Fig. 7.1 The spectrogram of source and synthesis speech of each syllable

표 2. 청취율[3]
Table 2. The recognition accuracy

청취율	대상단어					
	누	춤	겨자	타파	자동차	우리말
	5/10	8/10	7/10	9/10	7/10	3/10
	65%					

그림 7.1은 6개의 합성음과 원음성의 스펙트로그램이다. 합성에서 천이구간(F2) 처리에 대한 두가지 방법이 있다. 그림 7.1(a)의 /우/와 (d)의 두 번째 모음 /아/와 같이 2중모음 자체를 이용한 방법과 그림 7.1(d)중 첫 음절 모음 /여/와 그림 7.1(c)의 첫 음절모음 /아/의 뒷부분을 조정하는 LSP를 이용하는 방법이 있다. 유성 자음은 모음과 결합하는 경우에 서로의 피치에 맞도록 서로 조절하는데 그림 7.1(b)의 /우/와 /ㅁ/의 결합과 그림 7.1(e)의 두 번째 음절에서 /오/와 종성자음 /ㅇ/, 그림 7.1(f)중 마지막 음절 /아/와 /ㄹ/이 있다.

F2를 조절해야하는 천이구간에서는 /아, 야, 어, 여/의 경우는 쉽게 구현할 수 있지만, /오, 우/에서는 /아, 야, 아, 여/의 모음처럼 F1과 F2가 1kHz이하에 존재하지 않고 F2가 2500Hz 이상에 존재하면서 그 에너지가 약해 조절하기가 용이하지 않다.

Ⅵ. 결 론

본 연구에서 주안점을 두었던 부분은 음성합성에서 많은 문제가 되는 막대한 데이터 양과 천이구간의 처리이다. 데이터가 많이 필요한 이유는 음운환경에 따라 달라지는 서로 다른 정보를 처리하기 곤란하기 때문이다. 이런 이유로 처음부터 각각 환경이 다른 데이터들을 수집하게 되고, 수집한 데이터를 알맞게 재구성함으로써 음성을 합성한다. 하지만 앞서 기술한 것과 같이 많은 메모리와 각 환경에 따른 데이터를 수집해야 하는 어려움이 존재한다. 이러한 단점을 극복하고자 음소를 이용하였지만, 음소의 특성상 음운환경에 대한 적응성이 부족하였다. 적응성이라 한다면 음과 음사이에서 발생하는 천이구간처리 문제이다. 단음의 음소데이터를 수집하였기 때문에 천이구간에 대한 정보를 가지고 있지 않다. 하지만 LSP line들을 통해서 일정수준의 포먼트의 조정이 가능하게 되었다. 분석을 통해 얻어진 LSP line들은 실험을 통해서 pair를 찾아내고, 찾아낸 pair를 조정함으로써 변동시키고자 하는 포먼트를 조절하였다. 합성단위는 /이/ 모음의 초성 /ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ/ 13개와 중성의 단모음 /아, 어, 오, 우, 으 ,이 에/ 7개, 이중모음 /야, 여, 요, 유, 예, 와, 워, 왜, 의/ 9개, 종성 /ㄴ, ㄹ, ㅁ, ㅇ/ 4개로 총 33개의 음소를 이용하여 유음이나 모음등 유성음의 합성에 PSOLA방식을 채택하였다.

문제점으로는 첫째, LSP의 적용시 각 line이 겹치거나 간격의 변화에서 오는 주파수 왜곡현상으로 고주파대역에서의 조절과 주파수에너지의 증감이 쉽지 않으며, 종성폐쇄음중에서 /ㄱ, ㅂ/에 대한 구현이 어려웠다. 둘째, 경음은 이에 대한 정보와 분석이 부족하여 구현하지 못하였다. 셋째, 비슷한 음가를 가지고 있는 파열음과 파찰음에 대한 명료도가 불분명하며, 넷째로 음운환경에 따른 정확하고 풍부한 음의 장단에 대한 정보가 부족하며, 음절이 늘어나면서 중간에 나타나는 모음의 길이가 줄어

드는 현상에 대한 명료도를 유지하기가 쉽지 않다.

이런 문제점들을 해결하기 위해서는 자음과 모음의 결합에서 발생하는 펄스와 같은 잡음의 제거와 포먼트 조정이외에 특정지역의 주파수 에너지의 자유로운 조절을 더 연구해야 한다.

참 고 문 헌

1. 박애희, 양진우, 김순협, "음소단위를 이용한 소규모 문자-음성변환 시스템의 설계 및 구현", 한국음향 학회지 제14권 제3호, pp. 49-60, 1995.
2. 권혁제, 이태진, 김종교, "자음의 주파수 변조를 통한 소규모 음소 PSOLA 음성합성", 하계종합학술대회 논문집, pp. 653-656, 1996.
3. 권혁제, 최형기, 김종교, "음소 합성을 위한 음의 전이구간 처리", 대한전자공학회 추계종합학술대회논문집, 1996.
4. Lawrence Rabiner, Biing-Hwang Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.
5. Shuzo Saito, Kazuo Nakata, *Fundamentals of Speech Signal Processing*, ACADEMIC PRESS, pp. 126-132, 1985.
6. A. M Kondoz, *Digital speech coding for low bit rate communication systems*, JOHN WILEY & SONS, pp. 79-115, 1995.
7. Hideyuki Mizuno, Masanobu Abe, Tomohisa Horokawa, "Waveform-Based Speech Synthesis Approach with a Formant Frequency Modification," *ICASSP*, pp. II-195~II-198, 1993.
8. 김상훈, 지민제, 최도현, 한희일, "글소리 II에서의 신호처리," 제1회 ETRI 음성, 언어 및 음향정보처리 워크샵 논문집(1993. 4), pp. 91-96, 1993.
9. 정국, 구희산, 이찬도, 김종미, 한선희, "음성인식/합성을 위한 국어의 음성음운론적 특성 연구," 한국음향학회지 제 13권 16 호, pp. 31-43, 1994.
10. Yoshinori Sagisaka, "Speech Synthesis from Text," *IEEE Communications Magazine*, pp. 35-41, Jan. 1990.
11. Nobuyuki Katae, Tatsuro Matsumoto, Shinta Kimura, "High-Quality Japanese Text-To-Speech System: NARSYS," *EUROSPEECH 95*, pp. 1861-1864, 1995.
12. Christan HAMON, Eric MOULINES, Francis CHARPENTIER, "A Diphone synthesis system based on Time-domain Prosodic Modifications of Speech," *ICASSP 89*, pp. 238-241, 1989.
13. 하정호, 정재호, "합성음 구현을 위한 음의 억양과 장단변환 연구," 제11회 음성통신 및 신호처리 워크샵 논문집, pp. 328-333, 1994.
14. John R. Deller, Jr., John G. Proakis, John H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Publishing Company, pp. 331-333, 1993.
15. Panos E. Papamichalis, Ph. D., Practical Approaches to Speech Coding, Prentice-Hall, Inc., pp. 143-145, 1987.
16. Shuzo Saito, *Speech Science and Technology*, IOS Press, pp. 79-90, 1992.
17. Sadaoki Furui, *Digital Speech Processing, Synthesis, and Recognition*, Marcel Dekker, Inc., pp. 126-137, 1991.

▲권 혁 제(Hyuck-Je Kwon) 1972년 1월 12일생

1997년 2월 : 전북대학교 대학원 전자공학과 졸업(공학석사)

1997년 3월~현재 : 전북대학교 대학원 전자공학과 박사과정

※주관심분야 : 음성인식, 합성

▲조 순 계(Soon-Kye Cho) 1958년 1월 27일생

1984년 2월 : 원광대학교 전자공학과 (공학사)

1988년 2월 : 숭실대학교 대학원 전자공학과(공학석사)

1994년 3월~현재 : 전북대학교 대학원 전자공학과 박사과정

1988년 3월~1990년 2월 : 한국과학기술원 과학기술대학 조교

1990년 3월~현재 : 조선대학교 공업전문대학교 전자통신과 부교수

※주관심분야 : 음성부호화, 영상압축, 디지털 통신

▲김 종 교(Chong-Kyo Kim)

현재 : 전북대학교 전자공학과 교수

한국음향학회지 제17권 1호 참조