E-025

乱流を音源として含む音声の分析・合成・知覚
Turbulant Source Speech Analysis, Synthesis, and Perception

伏木田　勝信†
Katsunobu Fushikida

１．まえがき

雑音成分は自然性を知覚させるために重要な役割を担っていると考えられる。音声においては乱流音源がその他の音響パラメータのランダムなゆらぎ等とともに自然性を感じさせる要因となっている。

乱流を音源として持つ無声子音や囁き音声に関する研究としては従来、生理学的な検討 [1,2,3] 、音響分析や合成的検討 [11,12]、乱流に関する物理的検討 [4,5] など様々な検討が行われている。

音声の産出過程は Fig. 1 に示されるように肺から送出される気流が声道と呼ばれる音響管を通過する際に変調され口から放射されるという音響エネルギーの輸送現象とみることができる。この際、横隔膜や喉頭、舌などの調音器官の動きの生理学的な制御も重要となる。

通常の呼吸時におけるように肺からの呼気（気流）が強くなく、途中に狭めがない場合は層流としてエネルギーが輸送されるが、気流の速度が大きく、舌と口蓋による狭めや、声門部における狭めや声帯振動がある場合は、角運動エネルギーを持つ渦から構成される乱流を発生させることにより大きいエネルギーを輸送させていると考えられる。この際、乱流によって生じる音波は無声子音や囁き音声の音源となっている。

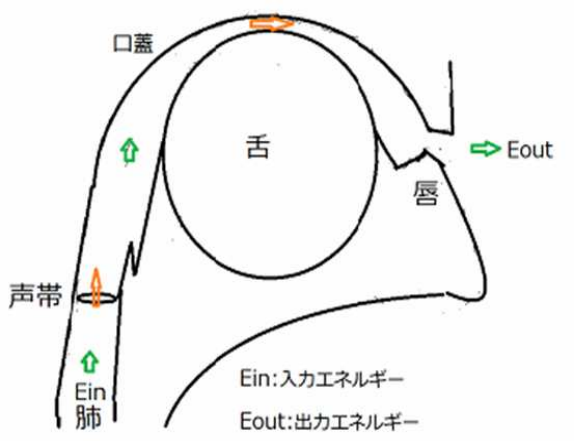


Fig. 1 Schematic diagram of speech production.

乱流を構成する渦の角速度の分布は、エネルギーの効率的な輸送行うために、音響管の周波数特性の影響を受けていると考えるのが妥当であろう。

人間の音声産出モデルの構築は、自然性が高く多様な音質を持つ人工音声の生成に際して重要なものとなる。

この論文では、①ＡｂＳ的フォルマント抽出方式、②乱流音源を含む音声のフォルマント等の分析結果、③乱流音源モデルに基づく音声合成実験と試聴テスト、知覚的ピッチ検出、④ピッチとフォルマントの相関についての分析および合成的検討結果について述べる。

2-1．フォルマント抽出方式

フォルマントの抽出の際には抽出を行う周波数帯域内に含まれるフォルマント数（NOF: Number Of Formants）を考慮する必要がある。NOF は声道長と対応しており性別、音韻、調音様式などによって変化するものである。

ここでは MSF(Multi-Step-Focusing)-ADIF(Autocorrelation-Domain-Inverse-Filtering) による AbS (Analysis-by-Synthesis) 的フォルマント抽出[6,9]とダイナミックプログラミング（DP）による NOF の最適選択方式 [7] を用いた。[15]

2 次の ADIF は、入力、出力の自己相関値をそれぞれ、r, v とすると 2 次の線形予測係数 $\{\alpha_1, \alpha_2\}$ より次の (1),(2) 式により効率よく算出されることが知られている。

$$r_i = (1+\alpha_1^2+\alpha_2^2)v_i - (\alpha_1-\alpha_1\alpha_2)(v_{i-1}+v_{i+1}) - \alpha_2(v_{i-2}+v_{i+2}) \quad (1)$$

ここで波形のサンプル値を S_n とすると
である。

$$v_{i-1} = \sum_{n=1}^{N} S_{n-1}S_{n-i} = \sum_{n=1}^{N} S_{n-2}S_{n-i-1}$$

また、Fig. 2 に示されるように、フォルマント数個分の 2 次の ADIF を縦列に接続し予測残差を算出比較して最適なフォルマントの組み合わせを粗い推定から詳細な推定へと focus して推定することにより安定した抽出を行なうことができる。

ここでは、フォルマント数（NOF）の推定方式として、前記予測残差パワーを最小とする NOF を DP により最適選択する方式 を用いている。

Fig. 3 に男声 /sasis(u)seso/（/u/ は無声化している）のフォルマント分析結果例を予測残差パワー値とともに示す。

フォルマント抽出の際に得られる予測残差（prediction residual）は乱流の発生強度と相関があると推測される。

この全極型フィルターの残差波形情報は音節明瞭性の改善に寄与することが 報告されている[13] 。

†KF 研究所, KF Laboratory

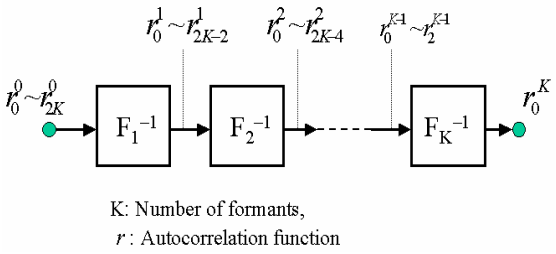
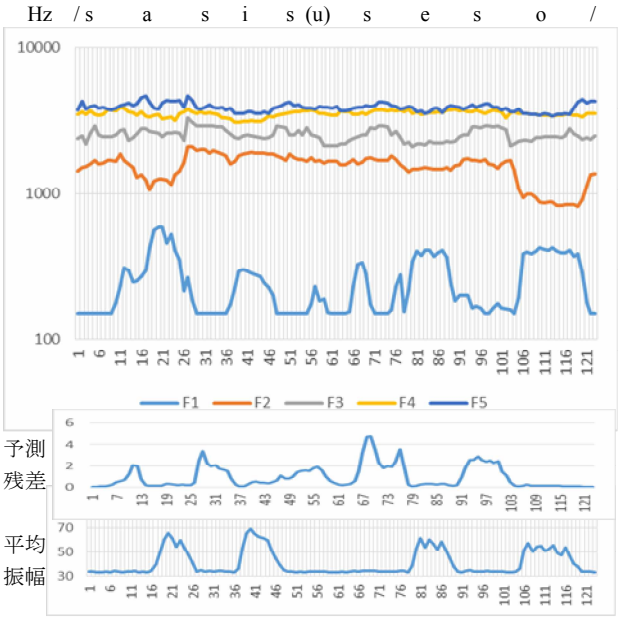


Fig. 2. Cascaded Second order ADIF.



(Horizontal axis: analysis frame number)

Fig. 3. An example of formant analysis result.
F6 以上のフォルマントは表示されていない

2 -2.無声（囁き）母音、無声子音の分析

音声サンプルとしては男女各 1 名が発声した日本語の有声 5 母音（ /a,i,u,e,o/ ）、および無声母音。無声摩擦音 /s/, /h/ を含む音節連鎖を用いた。サンプリング周波数は 22.05kHz 、フレーム周期は 200 sample points （約 9.1msec）, 分析ウィンドー長は 512 sample points の Hamming 窓を用いた。

フォルマントの先鋭度を表す Q 値としては Q =10 を一貫して用いた。

2 -3. フォルマント分析結果

通常の有声母音、無声母音（囁き母音）、無声子音について分析したフォルマント周波数および予測残差（residue）の結果例を Table 1, 2, 3 に示す。分析対象サンプルは男声 1 名である。

Table 3 は無声子音（摩擦音 /s/, /h/）のフォルマント、予測残差の分析結果例を示す。

Table 1. Formant and residual data of male whispered vowels

whisp.	F1	F2	F3	F4	F5	resid.
a	851	1285	2730	3358	4240	0.919
i	421	2277	3075	3592	4703	3.457
u	503	1152	2455	3222	3873	1.236
e	725	1929	2559	3715	5451	2.065
o	611	896	2943	3307	4218	0.902

Table 2. Formant and residual data of male modal vowels

modal	F1	F2	F3	F4	F5	resid.
a	612	1151	2892	3258	3708	0.637
i	291	2333	3283	3566	5509	0.669
u	307	1330	2463	3521	4632	0.473
e	451	1645	2363	3539	3651	0.870
o	381	861	2071	3038	3651	0.384

Table 3. Formant and residual data of fricative consonants

	F1	F2	F3	F4	F5	F6	F7
	F8	F9	F10	F11	F12	F13	res.
s(i)	149	2011	2905	3555	3933	5189	6140
	6881	7304	7692	9389	9389	10387	2.070
s(o)	167	1713	2875	3735	3913	5243	5861
	7044	7534	8733	8962	9438	10996	2.895
h(i)	157	1767	3020	3630	4374	5122	6269
	6971	8143	8228	9786	10996	*	0.458
h(o)	386	1020	2500	3687	3687	5311	5325
	6467	7191	8378	8620	9837	10826	0.369

＊NOF は/h(i)/ は 12 で、それ以外は 13 と最適選択された。( ) 内は後続母音を表す

囁き母音（whispered vowel）のフォルマント周波数は通常の（modal）発声の母音のフォルマント周波数よりも高い傾向があり、この結果は従来の結果 [9]とも一致している。

乱流音源の強度と予測残差パワーには相関が認められ、摩擦音の予測残差パワーは母音等より大きい。また、母音でも /o/,/u/ は小さい傾向が認められ、鼻音などでは予測残差パワーは比較的小さい傾向がある[7,8]。

3．音声合成 実験と結果

2-3 の分析結果に基づいて囁き母音（無声母音）と無声摩擦音の合成実験を行った。

3 -1. 乱流音源モデル

乱流は様々な角運動量を持つ渦から構成され乱流音源の周波数スペクトルはそれらの渦の分布を反映している。

ここでは乱流音源の周波数スペクトルは声道の共鳴特性と一致すると仮定した。これは、エネルギーの効率的輸送という観点からも合理的であると考えられる。

乱流音源モデルとしては、Fig,4 に示されるように声道共鳴と同様の周波数スペクトルを持つ方式を用いた。なお、このモデルは口笛の合成に用いられたものと同様の方式である [10] 。

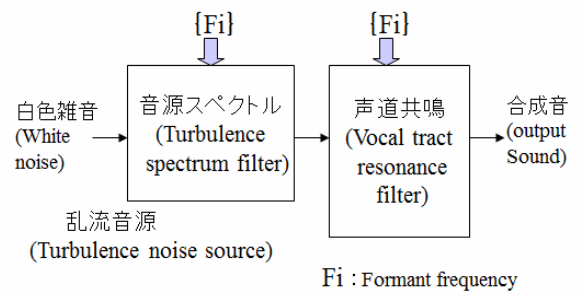


Fig. 4. Synthesis filter model with turbulence source.

提案する乱流音源モデルを用いた Fig. 4 の音声合成モデルの周波数スペクトルを、白色雑音音源を用いた場合と比較して Fig. 5 に示す。提案モデルではフォルマントピークが強調されるとともにディップも強調されている。

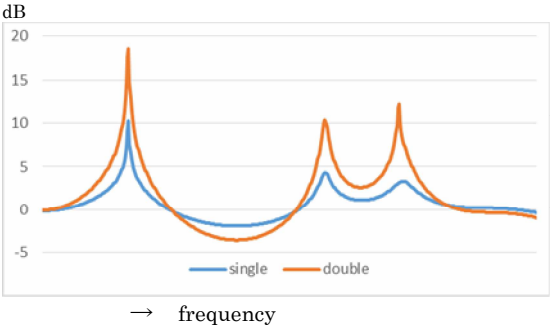


Fig. 5. Power spectrum of two kinds of synthesis filter
① Single: White noise source
② Double: Proposed turbulent source

3-2．無声母音（囁き音）の合成

通常の（modal）母音と囁き(whisper)母音の分析の結果得られたフォルマント周波数（Table 1,2）を用いて無声（囁き）母音(/a, i, u, e, o/)の合成実験を行った。

従来の白色雑音を音源とした場合と 3 -1 の乱流音源を用いた場合の合成音とを対比較で試聴した。その結果、上記乱流音源モデルにより作成された合成音の方が透明性が高く自然であることがわかった。

また、音韻の明瞭性に関しては Table 1 よりも通常の有声母音分析データ (Table 2) の方が高いと判断された。

さらに乱流音源のフィルターを縦列接続して合成音を作成したところ、透明性が増大することが認められた。

囁き母音のみならず通常の有声母音においても乱流音源が存在し母音の種類によってその強度が変化しており、予測残差とも相関があると考えられる。

3-3．無声子音の合成

無声音節の分析結果（Table-3）に基づいて無声摩擦音 /s/ と /h/ について、白色雑音を音源とした場合と 3 -1 の乱流音源モデルを用いた場合と合成音サンプルを作成した。

対比較の試聴テストを行った結果では、後者のほうが透明感があり自然性に優れていた。なお、原音声スペクトルの近似度の向上策としては、極零フィルターによる残差スペクトルの近似方式も報告されている[8]。

4. フォルマントとピッチの相関分析

ピッチとフォルマントの相関を明らかにするために有声音連鎖と母音の音階発声音の分析を行った。Fig. 6 に男声サンプル/aoiyane/ の第 1 フォルマント（F1）とピッチ周波数（f0）の倍音の分析結果例を示す。母音定常部でピッチ周波数（基本周波数 f0 とその倍音 n*f0）とフォルマント周波数（F1）との相関（一致）がみられる。

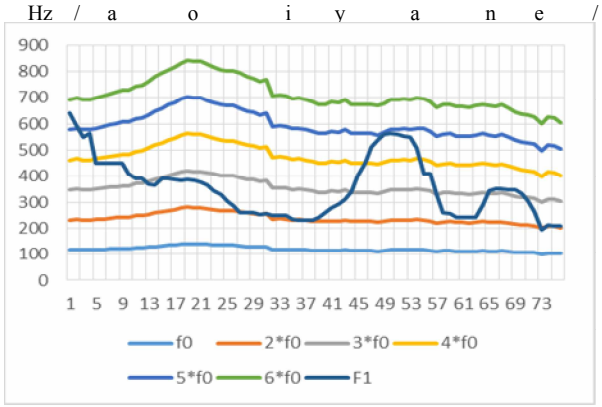


Fig.6. Analysis example of pitch and formant frequency

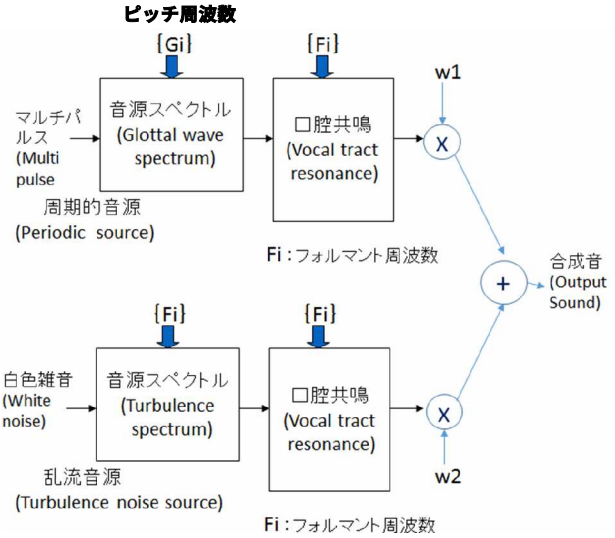
Table 4 に母音 /i/ の場合の 1 オクターブの音階発声音の分析結果例を示す。F1 は f0 の約 2 倍になっている。高い音階 si と do では NOF は 11 に減っている。

Table 4 Formant (F1,F2) and pitch (f0) analysis example

/i/	NOF	res.	f0	F1	F2
do	12	0.226	126.7	234	1851
re	12	0.268	123.9	258	1804
mi	12	0.193	132	265	1914
fa	12	0.260	135.3	261	1704
so	12	0.295	156.4	291	1842
ra	12	0.409	176.4	336	1758
si	11	0.427	198.6	393	2016
do	11	0.359	196.9	391	2146

5 有声音と無声音の重畳による合成実験

Fig. 7 に示されるように、周期的音源を用いた合成音波形と乱流音源を用いた合成音波形を重畳する音声合成モデルにより定常母音の合成実験を行った。



w1,w2 : weighting coefficient

Fig. 7 Speech synthesis model with glottal source and turbulence source

乱流音源を重畳することにより合成音の自然性の向上が認められ、重み係数 w1,w2 を変化させることにより透明度や嗄声度の制御ができることが確かめられた。

５母音に対して１オクターブの音階を持つ定常音声を Table 4 の結果に基づいて周期的音源と乱流音源で作成し試聴した。その結果、乱流音源の場合でも音程がある程度知覚できることが分かった。

フォルマント周波数（F1, F2 ,F3）を音階周波数の整数倍に近くなるように音韻性の許容範囲内で調整して合成音を作成し、予備的な試聴実験を行った。その結果、調整しない場合に比べ、出力パワーが大きく透明性が増し、音階が識別し易いと判断された。

Fig.8 に乱流音源のみ (w1=0, w2=1) の無声合成母音 /e/ に対して、知覚モデルに基づいたピッチ検出方式[14]を用いてピッチ検出を行った結果例を示す。聴覚的にも５母音の試聴テストにより１オクターブの音階が知覚されることを確かめた。なお、口笛も声道中で生成される乱流音源で音階知覚が得られる現象例である[10]。

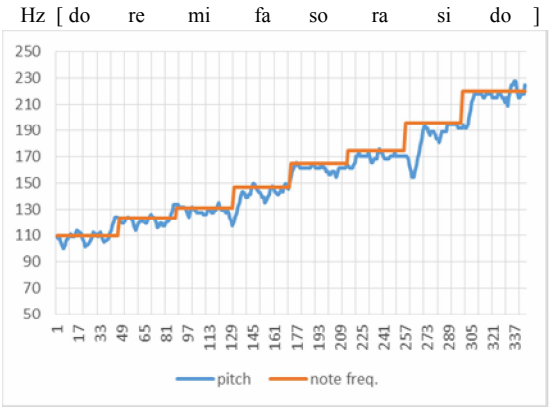


Fig. 8 Pitch extraction (by [14]) example for 1-octave note synthesized whispered vowel (/e/) using turbulence source

乱流音源の周波数スペクトルを声道の周波数スペクトルで近似する仮設は聴覚的には有効であると考えられる。なお、口笛の合成実験によっても透明性の改善に対して有効であることが確かめられている [10]。

フォルマントとピッチのチューニングは大きな音声のエネルギーの輸送を可能とする。歌声などにおける音質の調整方式としても効果的であると考えられる。

6．あとがき

乱流を音源として持つ無声子音や囁き音声のフォルマント分析結果に基づいて合成音の生成モデルを提案し、合成音の試聴テストを行った。その結果、乱流を音源とする音声の生成には乱流音源の生成モデルを用いて透明度の高い囁き母音等の無声音声が生成されることを明らかにし、乱流音源のみの音源（囁き音）でも音階が知覚される母音が合成できることを明らかにした。

また、フォルマントとピッチをチューニングする合成方式により有声音の透明性を高めることができることを示すとともに、乱流音源を持つ合成波形を重みを付けて重畳することにより自然性を高め、嗄声度を効果的に制御できることを確かめた。

参考文献

[1] Edited by S. Saito, *Speech Science and Technology*, (Ohmsha, Ltd., 1.4 Articulatory Control of the Larynx: Fiberscopic and EMG Studies, by M. Sawashima, pp.32-42,1992.
[2] K. Tsunoda, Y. Ohta, Y. Soda, Y. Niimi, and H. Hirose, Laryngeal adjustment in whispering. *Annals of Otology, Rhinology & Laryngology*, Vol. 106, pp.41-43, 1997.
[3] R. S. Weitzman, M. Sawashima, H. Hirose, “De-voiced and whispered vowels in Japanese," *Annual Bulletin, Research Institute of Logopedics and Phoniatrics*, vol. 10, 1976.
[4] 神部,ながれ 20,流れと音の物理, pp. 174-186, 2001.
[5] 大谷,半原,「非可聴つぶやき声発生時の声門流の数値シミュレーション」, 日本音響学会講演論文集, 2008-3.
[6] 伏木田,「自己相関領域で逆フィルタリングを用いたホルマントの多段推定方式」,日本音響学会,音声研究会資料、S81-41, pp. 323-329, 1981-10.
[7] 伏木田,「ＤＰを用いたフォルマント数の最適選択方式」、電子情報通信学会総合大会、D-14-25,2006.
[8] 伏木田,「音声の全極型予測残差の極零対フィルターによる分析」,情報処理学会第 72 回全国大会,2D-5,2010.
[9] K. Fushikida, ”A formant extraction method using autocorrelation domain inverse filtering and focusing method” , *Proc. IEEE ICASSP*, E2.8:2260-2263, 1988.
[10] 伏木田, 「フォルマントパラメータを用いた口笛の分析と合成」, 情報処理学会第 70 回全国大会講演論文集, 2-59-2-60, 2008.
[11] 松田 , 粕谷 , 「ささやき声の音響理論」 日本音響学会研究発表会講演論文集. I. 299-300, 1997-10
[12] Higashikawa, M, and Minifie, F. D., “Acoustical-Perceptual Correlates of “Whisper Pitch” in Synthetically Generated Vowels”, *J. Speech, Language, and Hearing Research*, Vol. 42, pp. 583-591,1999.
[13] 伏木田,落合,「素片編集型音声合成を目的とした自動分析合成の一方式」 日本音響学会,音声研資,S74-23,1974.
[14] 伏木田,「聴覚知覚特性に基づいたピッチ検出方式」, 第 11 回科学技術フォーラム FIT2012, E-029 , 2012.
[15] www5b.biglobe.ne.jp/~hfyfkf/Laboratory/Laboratory.htm