J-028

位相情報を操作した合成音声の脳波による評価と考察
Evaluation of the Synthesized Speech by Using Mismatch Negativity (II)

横野 和也† 吉田 秀樹†
Kazuya Yokono Hydecky Yoshda

1．まえがき

音声信号はヒトが発話している様に自然な抑揚で合成するのも困難であれば、逆に聴き取れなくなるまで完全に破壊するのも難しい構造をした音色であると云えよう。我々は音響構造の理解を深める上で、スペクトルの代わりに、時系列波形の極大値と極小値に着目してきた。極値の持つ情報（位相と振幅）は、音の高さ、大きさ、音色に対して本質的な働きをしていることが提唱されており[1]、極値を計測しても許容誤差を上回れば、合成できる音質は劣化し、ざらついた感じを与えてしまう。これに対し、隣り合う極値間の欠落した情報は、後からかなり自由に補うことが許されており、この緩やかな制約下での合成に伴う歪み成分は、知覚上問題にはならない。この意味で、音声信号の頑健な構造を理解する上で、極値からの切り口が期待されている。

実際に複合音の極値を計測するには、フィルターバンクを使用して複数の狭帯域チャンネルに分割して、うなり様の時系列波形にまで簡略化する前処理が必要となる。音響／音声合成には、極値間の正弦波補間と、チャンネル間の重ね合わせ処理も必要となる。煩雑ともとれる前後処理をしてまでも、極値の情報が重要となることは、神経生理学の研究結果から支持されるものである[2, 3]。入力音波は鼓膜と小耳骨の働きにより機械的な振動に変換されて、内耳にある蝸牛へと伝えられる。蝸牛内部の基底膜の振動には周波数特性が報告されており、特定の部位に配列した有毛細胞により検出される。音波は聴神経の内部では、スパイク波の時系列として情報表現されているが、その電荷の起源は、内有毛細胞の場合では不動毛の一方向へのたわみ、即ち過分極ではなく、脱分極にあると理解されている[4, 5]。これは内有毛細胞には謂ば検波特性があり、聴神経での短時間平均発火頻度が、濾波波形の極大値あるいは極小値の時刻でピークに達することと一致するからである。

本研究の目的は、極値の位相情報を変化させた合成音声の音質を、誘発脳波を計測することで客観的に評価することである。これにより極値の計測に際して許容誤差の範囲が見積もれる様になり、音響構造の中でも特に音色の理解に資することが期待される。脳波は近年、優れた識別器の提案に伴い、マンマシンインターフェースとしても利用の進んだ生体信号である[6-11]。我々は誘発脳波の中でも、ミスマッチ陰性(MMN) 応答と呼ばれる成分に着目した。ミスマッチ応答は、繰り返し呈示される刺激音列の中に僅かな変化が含まれていれば[12-15]、自動的に音質の差異を検出する精神作用を反映している[16, 17]。ミスマッチ応答は、被験者の意欲や関心、集中力に影響を受けることのない、有効な指標として活用が望まれる。

†北見工業大学大学院 情報システム工学専攻

2．計測方法

女性１名を含む健常被験者 10 名（年齢 19.1±0.3 歳、全員右利き）がボランティアとして実験に参加した。実験は非侵襲的に実施され、計測脳波は公表とし、計測中でも被験者の意思で実験は随時終了できる旨の充分な説明がなされた。被験者は簡易の静電対策を施した防音室 MC-3（155 x 255 x 210 cm, Music Cabin Co., Ltd.）の中で安静に椅子に座して読書をしながら、ヘッドホンを着用して刺激音を聞き流した。計測中は刺激音を無視させ、考え事も自由とした。刺激音は高頻度刺激と低頻度刺激から成るランダムな刺激列であり、両者の呈示割合は 6:1、音の長さは共に 144 ms、刺激間間隔 500ms とした。高頻度刺激は音節/ki/であり、44.1 kHz でサンプリングして、80-5,120 Hz に帯域制限したのに対し、低頻度刺激には位相情報をランダムに変化させた合成音声とした。先行研究[1]によれば、入力音声の有する周波数情報を狭帯域に制限することにより、濾波された波形から極大値と極小値の情報が抽出できる様になる。隣り合う極値間を正弦波様に後から補間することで、音声が合成できることが報告された。同手法に従って、80-5,120 Hz の帯域を１オクターブ毎に６分割して６個のチャンネルとし、それぞれのチャンネルから極値を抽出した。尚、当該帯域中には、音声信号のピッチおよび第１、第２フォルマント成分を網羅するのに、充分な帯域を有している。

図１に任意のチャンネルでの濾波波形の模式図を破線で示す。白丸で示された極値は、時間方向に一定量だけ移動させ、操作後の極値を黒丸で表した。この時、遅れ位相にするか進み位相にするかはランダムとした。同様の操作を６個のチャンネルに含まれる全ての極値について実施した。これにより位相誤差は d/T で記述する。低頻度刺激としては、位相誤差が 0%、5%、6%、7%、および 8%となる様に５種類の刺激音を用意して、同じ被験者には計５課題の実験を課した。この様に波形操作は極値の情報のみで実現し、図１の実線に沿って欠落した情報を補った後、６個のチャンネルを重ね合わせて合成波形とした。刺激音の音圧は約 65 dB SPL で、被験者毎に適宜微調整した。

脳波は国際式 10-20 法に基づく C3（左側頭頂）と C4（右側頭頂）の部位で両耳朶を基準電極として計測し (Digital Bio-Amplifier System 5202, ノイズ < 0.5 μ Vrms, NR 社製)、0.53-30 Hz のバンドパスフィルター(3 dB down, 12 dB octave/slope)と 50 Hz のハムフィルターに通した。高頻度刺激と低頻度刺激のそれぞれに同期させて、刺激前 100 ms から刺激後 400 ms の脳波を 80 回、加算平均処理した。同時に計測した眼電図が 150 μ V を超えた場合には、当該期間の脳波を加算平均から除外した。脳波は非線形システムである脳が生成する電気信号であるが、MMN 振幅を測定する目的で慣例に従って、低頻度刺激に対する応答波形から、高頻度刺激に対する応答波形を引き算して示した。MMN 振幅について MANOVA（２要因の分散分析法、部

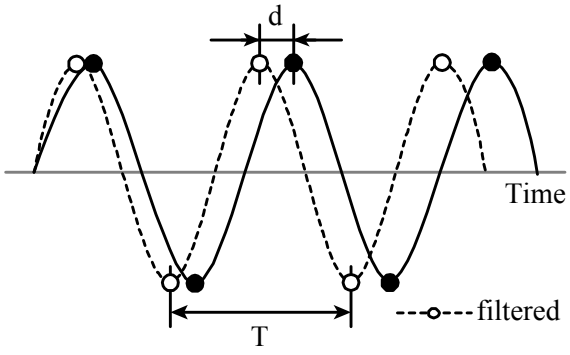

Fig.1 Phase error, d/T

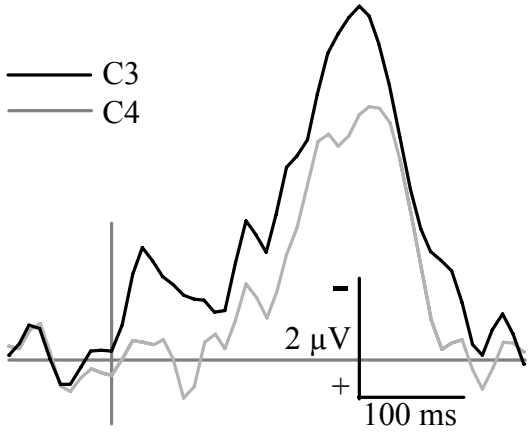

Fig.2 Differential waveforms

位{左側、右側}×実験課題 5 種類）を適用して有意水準 $p<0.01$ で下位検定（Tukey の方法）を実施した。

3. 結果

図２に左側と右側で計測された引算波形をそれぞれ実線と薄線で重ねて示す。両波形の振幅は、潜時約 245 ms で最大を示している。他の９名の被験者についても同様にして、MMN 振幅を計測した。

図３には５種類の課題について観察された MMN 振幅の平均値と標準偏差を示す。MMN 振幅は実験課題により有意に変化しており、$F_{(4, 36)}=13.6$、下位検定により MMN 振幅は、位相誤差が 7%と 8%の時に有意に増大していることが示されている（課題間の位相誤差が 0%-7%の時 $t_0=2.1$、0-8%時 $t_0=2.2$）。計測部位による差異と、$F_{(1, 9)}=2.6$、課題と計測部位の交互作用については、$F_{(4, 36)}=1.4$、影響が観察されていない。

4. 考察

極値サンプリングする際に必要となる狭帯域フィルタリング処理は、入力音の周波数が基底膜の場所として表現されていることに相当する。外有毛細胞の不動毛は、フィルタリングした入力音が極大値あるいは極小値の時刻で振り切れ電荷を生成するのに対して、内有毛細胞ではどちらか片方の極値の時刻で電荷を生成する。こうして生成された電荷は聴神経を伝わるスパイク波列となることから、音波の聴神経内部での情報表現に、極値の情報は密接に関わっていると考えられる。実際に哺乳類の聴神経では、入力音の 4 kHz 以下の周波数について、位相同期現象が観察されることが特徴となっており[18]、この現象が音源局在にかかる情報処理を実現する上で重要となる。例えば 1,500 Hz 以下の純音では、入力音の到来する方位角を推定するのに、僅か 50 μs の位相ずれ(ITD)を検出する能力があるとされる[19]。こうしてみると、極値の位相をランダムに組み替えると云う操作は、脳内での情報処理に混乱を来すことに他ならない。実際に僅か 7 %の位相誤差が、MMN 振幅（絶対値）の有意な増大となって観察されており、極値の振幅情報を操作した時と同様に、刺激音はざらざらした感じとして知覚されていた。本研究結果は主観評価（Scheffe の一対評価法）を使用した先行報告（位相の許容誤差は 4 %から 9 %の間）[20]と一致しており、指標に MMN を導

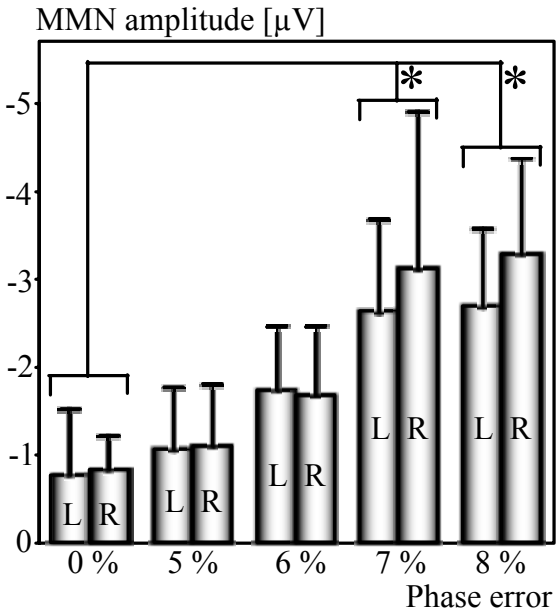

Fig.3 MMN amplitude, *p<.01

入することで、誤差の閾値をより詳細に見積ることに貢献した。

大脳の活動の非対称性については、言語処理時に右利きの被験者の大半について左側半球が優勢に活動するとした知見が[21]、様々な種類の脳機能イメージング装置により報告されている[22, 23]。これらの研究は、インピーダンスの低い脳脊髄液に遮られて充分な空間解像度を得ることの難しい、非侵襲な脳波の計測からは得られ難い観察となっている。音声／言語処理と云うのは、脳内の音響情報処理（連続情報処理）と音韻情報処理（範疇化情報処理）の両者を組み合わせた複合モデルにより説明されており[24]、脳内には音声信号を理解するための特有の機能が備わっていることが指摘されてきた[25]。前者の音響情報処理は、

入力音の物理的な特徴を自動識別するとしたミスマッチ応答に深く関連しているのに対して、後者の範疇化情報処理の存在は、刺激音に注意を傾けた状態で、後潜時(> 250 ms)の事象関連電位に非対称性が観察されることから支持されている[26, 27]。注意条件と云うのは、音声信号処理を実現する上で重要な役割を果たしていると考えられているが、脳波を使用した MMN 振幅には、半球間の非対称性は報告されておらず[28, 29]、これは我々の観察と一致している。

5. まとめ

極値の位相の許容誤差は 7 %未満であり、閾値は 6 %から 7 %の間に見積もられた。

参 考 文 献

[1] 入場健仁、吉田秀樹、藤原祥隆、岡田信一郎(2004):リアルタイム音響構造配信システムの開発 情報技術レターズ, LK-009, pp. 267-268.

[2] Dallos, P., Billone, M. C., Durrant, J.D. Wang, C.-Y., and Raynor, S. (1972): “Cochlear Inner and Outer Hair Cells: Functional Differences”, Science, Vol.177, pp. 356-358.

[3] Hudspeth, A. J. and Corey, D. P. (1977): “Sensitivity, Polarity and Conductance Change in the Response of Vertebrate Hair Cells to Controlled Mechanical Stimuli”, PNAS, Vol.74, pp. 2407-2411.

[4] Greenwood, D. D. (1990): “A Cochlear Frequency -Position Function for Several Species- 29 Years Later”, J. Acoust. Soc. Amer., Vol.87, pp. 2529-2605.

[5] Russel, I. J. and Sellick, P. M. (1978): “Intracellular Studies of Hair Cells in the Mammalian Cochlea”, Journal of Physiology, Vol.284, pp. 261-290.

[6] Obermaier, B., Muller, G. R., and Pfurtscheller, G. (2003): “Virtual Keyboard Controlled by Spontaneous EEG Activity”, IEEE Trans. Neural Sys. Rehab. Eng., Vol.11, No.4, pp. 422-426.

[7] Fabiani, G. E., McFarland, D. J., Wolpaw, J. R. and Pfurtscheller, G. (2004): “Conversion of EEG Activity into Cursor Movement by a Brain-Computer Interface (BCI)”, IEEE Trans. Neural Sys. Rehab. Eng., Vol.12, No.3, pp. 331-338.

[8] Jung, T.-P., Makeig, S., Stensmo, M. and Sejnowski, T. J. (1997): “Estimating Alertness from the EEG Power Spectrum”, IEEE Trans. Biomed. Eng., Vol.44, No.1, pp. 60-69.

[9] Anderson, C. W., Devulapalli, S. V. and Stolz, E. A. (1995): “Determining Mental State from EEG Signals Using Neural Networks”, Scientic Programming, Special Issue on Applications Analysis, Vol.4, No.3, pp. 171-183.

[10] Middendorf, M., McMillan, G., Calhoun, G. and Jones, K. S. (2000): “Brain-Computer Interfaces Based on the Steady-State Visual-Evoked Response”, IEEE Trans. Rehab. Eng., Vol.8, No.2, pp. 211-214.

[11] Suppes, P. and Han, B. (2000): “Brain-Wave Representation of Words by Superposition of a Few Sine Waves”, PNAS, Vol.97, No.15, pp. 8738-8743.

[12] Fitzgerald, P. G. and Picton, T. W. (1983): “Event-Related Potentials Recorded during the Discrimination of Improbable Stimuli”, Biological Psychology, Vol.17, pp. 241-276.

[13] Naatanen, R., Paavilainen, P., Alho, K., Reinikainen, K. and Sams, M. (1989): “Do Event-related Potentials Reveal the Mechanism of the Auditory Sensory Memory in the Human Brain?”, Neuroscience Letters, Vol.98, pp. 217-221.

[14] Naatanen, R., Paavilainen, P., and Reinikainen, K. (1989): “Do Event-related Potentials to Infrequent Decrements in Durations of Auditory Stimuli Demonstrate a Memory Trace in Man?”, Neuroscience Letters, Vol.107, pp. 237-242.

[15] Ford, J. M. and Hillyard, S. A. (1981): “Event Related Potentials (ERPs) to Interruptions of Steady Rhythm”, Psychophysiology, Vol.18, pp. 322-330.

[16] Naatanen, R., Gaillard, A. W. K. and Mantysalo, S. (1978): “Early Selective Attention Effect on Evoked Potential Reinterpreted”, Acta Psychologica, Vol.42, pp. 313-329.

[17] Naatanen, R. (1990): “The Role of Attention in Auditory Information Processing as Revealed by Event-Related Potentials and Other Measurements of Cognitive Functions”, Behav. Brain Sci., Vol.13, pp. 201-288.

[18] Palmer, A. R. and Russel, I. J. (1986): ”Phase-Locking in the Cochlear Nerve of the Guinea-Pig and It’s Relation to the Receptor Potential of Inner Hair-Cells”, Hearing Research, Vol.24, pp. 1-15.

[19] Makous, J. C. and Middlebrooks, J. C. (1990): “Two-Dimensional Sound Localization by Human Listeners”, J. Acoust. Soc. Amer., Vol.87, pp. 2188-2200.

[20] 吉田秀樹、角井健二、前田康成、藤原祥隆(2008): 極値サンプリング技術と許容誤差− wavファイルからの情報抽出− バイオメディカル・ファジイ・システム学会誌 Vol.10(2), pp. 123-131.

[21] Wada, J. and Rasmussen, T. (1960): “Intracarotid Injection of Sodium Amytal for the Lateralization of Cerebral Speech Dominance: Experimental and Clinical Observations”, J. Neurosurg., Vol.17, pp. 266-282.

[22] Thierry, G., Boulanouar, K., Kherif, F., Ranjeva, J.-P. and Demonet, J.-F. (1999): “Temporal Sorting of Neural Contents Underlying Phonological Processing”, NeuroReport, Vol,10, No.12, pp. 2599-2603.

[23] Shtyrov, Y., Kujala, T., Palva, S., Ilmoniemi, R. J. and Naatanen, R. (2000): ”Discrimination of Speech and of Complex Nonspeech Sounds of Different Temporal Structure in the Left and Right Cerebral Hemispheres”, NeuroImage, Vol.12, pp. 657-663.

[24] Klatt, D. E. (1982): “Speech Processing Strategies Based on Auditory Models”, In Carlson, R. & Granstrom, B. (Eds.) “The Representation of Speech in the Peripheral Auditory System”, Amsterdam: Elsevier.

[25] Whalen, D. H. and Liberman, A. M. (1987): “Speech Perception Takes Precedence over Nonspeech Perception”, Science, Vol.237, No.4811, pp. 169-171.

[26] Celsis, P., Doyon, B., Boulanouar, K., Pastor, J., Demonet, J.-F. and Nespoulous, J.-L. (1999): “ERP Correlates of Phoneme Perception in Speech and Sound Contexts”, NeuroReport, Vol.10, No.7, pp. 1523-1527.

[27] Szymanski, M. D., Yund, E. W. and Woods, D. L. (1999): “Human Brain Specialization for Phonetic Attention”, NeuroReport, Vol.10, No.7, pp. 1605-1608.

[28] Aaltonen, O., Niemi, P., Nyrke, T. and Tuhkanen, M. (1987): “Event-Related Brain Potentials and the Perception of a Phonetic Continuum”, Biological Psychology, Vol.24, pp. 197-207.

[29] Aulanko, R, Hari, R., Lounasmaa, O. V., Naatanen, R. and Sams, M. (1993): “Phonetic Invariance in the Human

Auditory Cortex”, NeuroReport, Vol.4, No.12, pp. 1356-1358.