2006－SLP－63 (3)
2006／10／20

ハイパスフィルタ付きNAMマイクロホンによるNAMの疑似ささやき声化

中島淑貴 鹿野清宏
奈良先端科学技術大学院大学 情報科学研究科

要旨：NAMマイクロホンにより収録されるNAMは，声質変換などの技術で，通常音声やささやき声に変換して無音声電話などの通信に使う方法があるが，NAMマイクロホン回路に，あるカットオフ周波数とスロープ特性をもつハイパスフィルタを組み込むことにより，出力は聴覚的に擬似的なささやき声様の音声となり，学習の必要がなく，ローコストでリソース消費のない通信利用が可能になる．今回我々は理想的なハイパスフィルタのカットオフ周波数とスロープを決めるために，HPF-NAMの聴覚的な評価実験を行った．

Whisper-Like Voice by NAM Microphone with High-Pass Filter

Yoshitaka Nakajima,, Kiyohiro Shikano,
Graduate School of Information Science, Nara Institute of Science and Technology (NAIST)

SUMMARY: Non-Audible Murmur (NAM) can be used as an input interface for confidential telecommunication that annoys nobody due to its conversion to normal speech or a whisper voice using the technology of statistical voice conversion, so-called "non-speech telephony." Instead of using statistical voice conversion we installed an analog high-pass filter only of a resistor and a condenser into the NAM microphone amplifier circuit, and converted NAMs to a whisper-like voice (HPF-NAM) at presumably the lowest resource cost. In this paper we perform perceptual evaluations of naturalness and intelligibility on HPF-NAMs to determine the optimal cut-off frequency and filter slope of the high-pass filter.

1. はじめに

非可聴つぶやき（Non-Audible Murmur: NAM）[1]は，そのセンサーであるNAMマイクロホンにより収録可能である，NAM音を増幅してそのまま聞くと，「こもったささやき声」のように聞こえるが，慣れればある程度聞き取ることができる[2]．また，GMMなどを用いた声質変換の技術を使い，通常音声に変換する技術もある[3]．しかし基本周波数の元来見られないささやき声に変換すれば，予測したF0パラメータを与えて通常音声に変換するよりも韻律に不自然性が聞き取れず，人間の発声した音声として，自然に聞こえることがわかった[4]．現時点で，あらかじめ学習を行えば，リアルタイムの変換も可能になっている．しかしこの方法は，学習が必要であったり，ソフトウェア的にリソースを消費したり，若干のディレイが起こる．

今回我々は，声質変換の技術を用いず，マイクアンプ回路に抵抗とコンデンサのみのハイパスフィルタを組み込み，非常に低コストで擬似的にささやき声に似た音声（以下：HPF-NAM）を出力するNAM

マイクロホンを試作した．デモンストレーションなどでの使用実感として，ささやき声に近い自然性と聞き取り率の改善が得られているが，NAM を人間の耳に聞き取りやすくするのに適したハイパスフィルタのカットオフ周波数とスロープ特性を調査するために，疑似ささやき声としての自然性と，人間の聞き取り認識精度の観点から，予備的な評価実験を行った．

2. HPF-NAM

Figure1 にサンプリングレート 8KHz の NAM と，アナログハイパスフィルタを組み込んだ NAM マイクロホンで収録した HPF-NAM（カットオフ周波数 1KHz，スロープ特性 6dB/oct），そして対照として気導音のささやき声のスペクトラムと，文頭母音「a」の周波数分析のグラフを掲げる．

数多くの NAM をリアルタイムでスペクトラムを観察しながら収録した経験上，NAM はスペクトル上低域の強調が著しく，750Hz 以下の音韻情報にあまり関係の無い部分があまりにも濃い印象を受けていた．NAM を多く聞いた経験でも、非常に低域に「ゴーッ」という定常的な乱流雑音の響きが常に乗っており、それが聞き取りの邪魔をしているのではないかと考えた．それに対して気導音であるささやき声を収録してスペクトラムを観察すると 750Hz 以下にほとんどフォルマントを認めない．ささやき声にないのであれば，この部分を弱めるか切って捨てたら、どんな音声になるかという発想からハイパスフィルタを使用した．

また NAM の信号波形は，気導音声に比して母音より子音の振幅が大きくなる傾向があるが[1]，高いカットオフ周波数のハイパスフィルタをかけるほど，子音の振幅が低くなり，信号波形の外見が気導音声に近づくことも経験された．

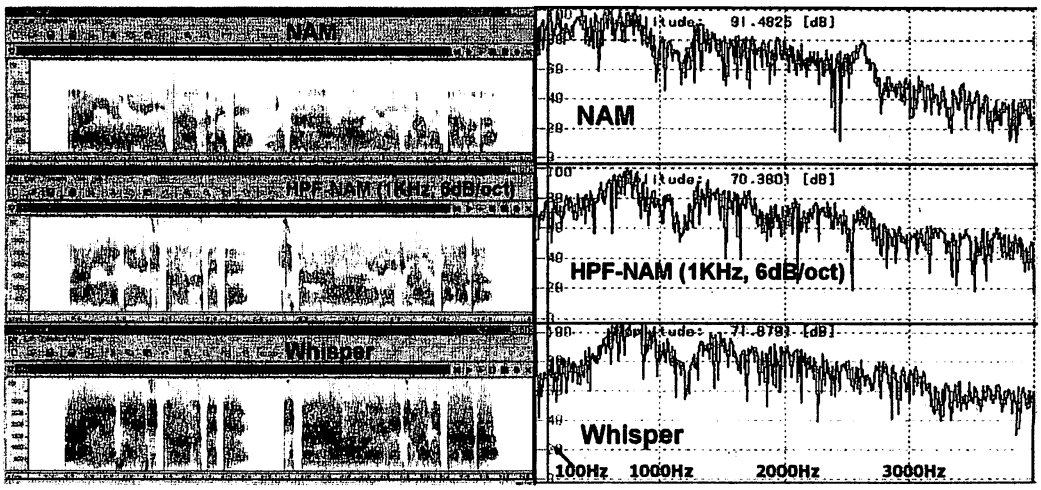

Figure1: NAM，HPF-NAM，ささやき声のスペクトラムと文頭「a」の周波数分析

3. 実験条件

自然性を評価するために，ミツミ製ソフトシリコーン型 NAM マイクロホンを用いて，「もしもし，こんにちは．お元気ですか」という発話内容の文章を 1 個収録した．また単語数 20～30 の新聞記事 8 文章と，6～7 音素の無意味単語 8 個を，男性一名が NAM 発話で読み上げて収録した．すべて量子化 16bit，サンプリングレート 8KHz で PCM 録音した．他に対照として，接話マイクを用いて，ささやき

声で読み上げ，同条件で録音した．尚，ささやき声はNAM発話より明らかにパワーが大きく，近くの人に聞こえるぐらいの音量で明瞭に発話した．

この収録したNAMの全17発話に対し，スロープ特性は「急峻なもの」と「なだらかなもの」の2種類，カットオフ周波数はそれぞれのスロープ特性につき200Hzきざみで200～2000Hzまでのソフトウェア的なハイパスフィルタをかけて処理し，1発話につき合計20個のサンプルを作成した．対照としてオリジナルNAMと気導音ささやき声を入れて1発話につき合計22個のサンプルとし，これらに対し，三つの聴覚的な実験を行った．使用したハイパスフィルタのスロープ特性は，Figure.2とFigure.3に示す通りである．

被験者は様々な業種の音声・音情報を専門分野としない，17歳から72歳までの，男性10名（平均年齢45.3歳），女性12名（平均年齢33.9歳）の計22名である（平均年齢39.1歳）．

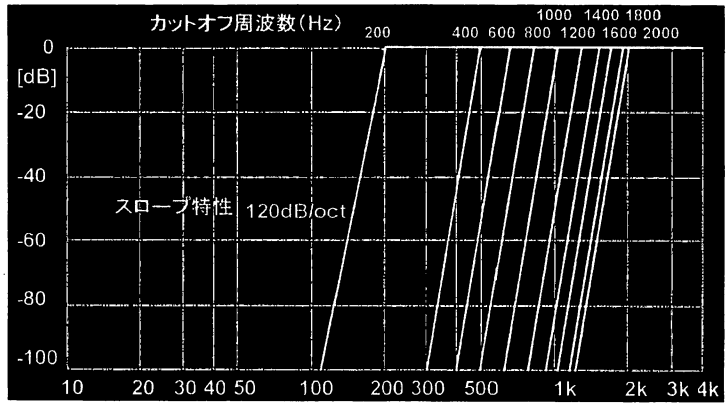

Figure2: フィルタのスロープ特性1 (SLOPE1)

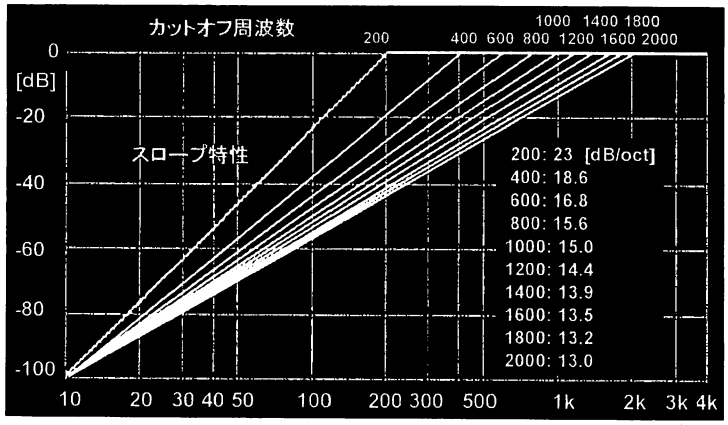

Figure3: フィルタのスロープ特性2 (SLOPE2)

4. 自然性に対する評価

まず疑似ささやき声としての自然性を評価するための実験を行った．対照のNAMと気導音ささやき声を含めた「もしもし，こんにちは．お元気ですか」という発話内容の全22サンプルにつき，総当り

制でペアにした．総計231のペアを21人の被験者に11ペアずつランダムに割り振った．密閉式ヘッドホンで二つの録音を一回聞いてもらい，二者択一で「自然なささやき声により近く聞こえる方」を選択してもらった．リーグ戦と同様，選択率（勝率）の高い方がより多くの人に「ささやき声としての自然性」が感じられたことになる．

結果をFigure.4に示す．8割を超える選択率を示したのは1000Hz (SLOPE2)で,6割を超えたものが急峻なSLOPE1の800Hzと1000Hz，なだらかなSLOPE2の800Hz, 1000Hz, 1200Hz, 1600Hzであった．各カットオフ周波数につきSLOPE2の方が200Hz以外は全体的に高い選択率を示し，NAMの選択率以上であった．急峻なSLOPE1では1400Hz以上で極端に選択率が低下する．対照のささやき声は，当然一番高い選択率を示したが，選択率は必ずしも10割ではなかった．

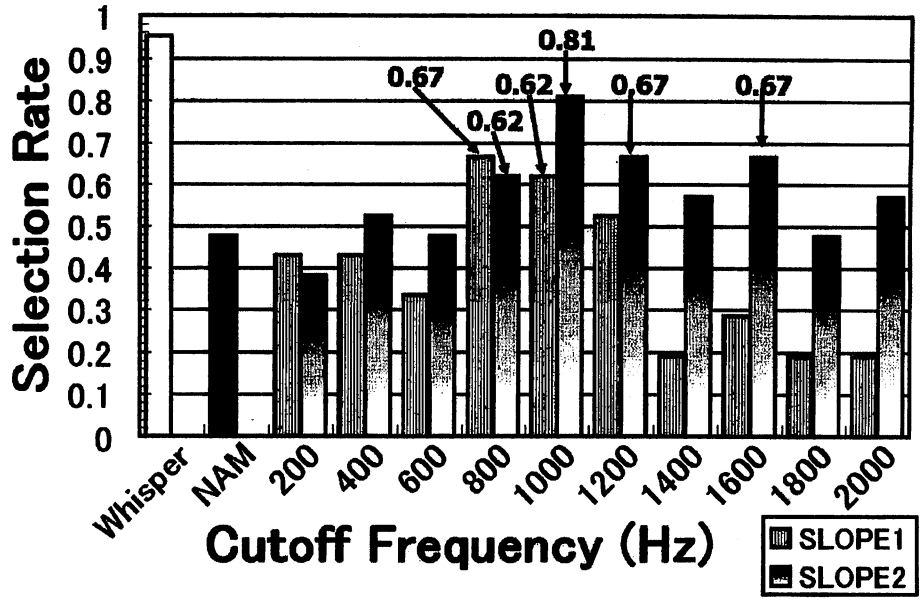

Figure 4: ささやき声としての自然性に関する評価

5. 人間の聞き取り認識精度

新聞記事8文章，無意味単語8個の計16問題のそれぞれにつき，NAMと気導音ささやき声の2対照を含めた全22種のサンプルを，22名の被験者に，問題ごとにランダムな順列で割り振った．密閉式ヘッドホンにて，各問題を繰り返し無制限に聞かせ，それを書き取ってもらった．新聞記事聞き取りの結果をFigure.5に示す．認識率は機械認識の場合と同様，単語認識精度で計算した．

対照のNAMの認識精度は0.849であったが，SLOPE1の200～600Hz，SLOPE2の200～1400HzでNAMと同等かそれ以上の単語認識精度を示した．特にSLOPE2の800Hz，SLOPE1の200Hzと400Hzでは5%以上の認識精度の上昇を見た．

無意味単語聞き取りの結果をFigure6に示す．未知語に対する聞き取りの精度として音素認識精度を使った，これは単語認識精度の単語を音素で置き換えたものである．

無意味単語でも全体的な傾向は新聞記事の場合と同じであったが，オリジナルNAMの音素認識精度がやや高かったため，それ以上の認識精度を示したものは，SLOPE2の200Hzと800Hzにとどまった．

興味深いのは対照の気導音ささやき声でも，音素に対する認識精度は7割程度であり，HPF-NAMの方が高い認識率を示すことがあった.

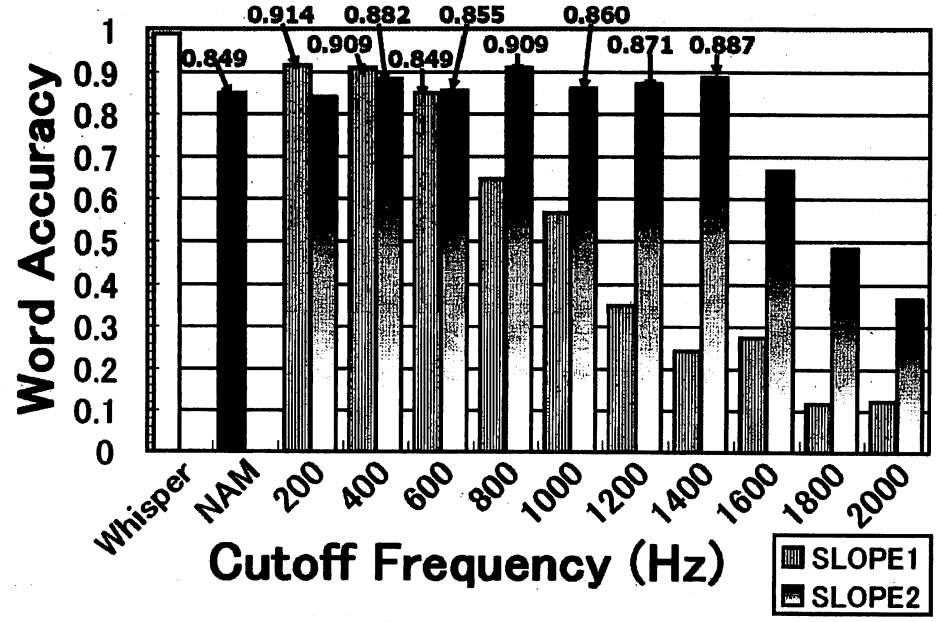

Figure 5: 人間の聞き取り認識精度（新聞記事）

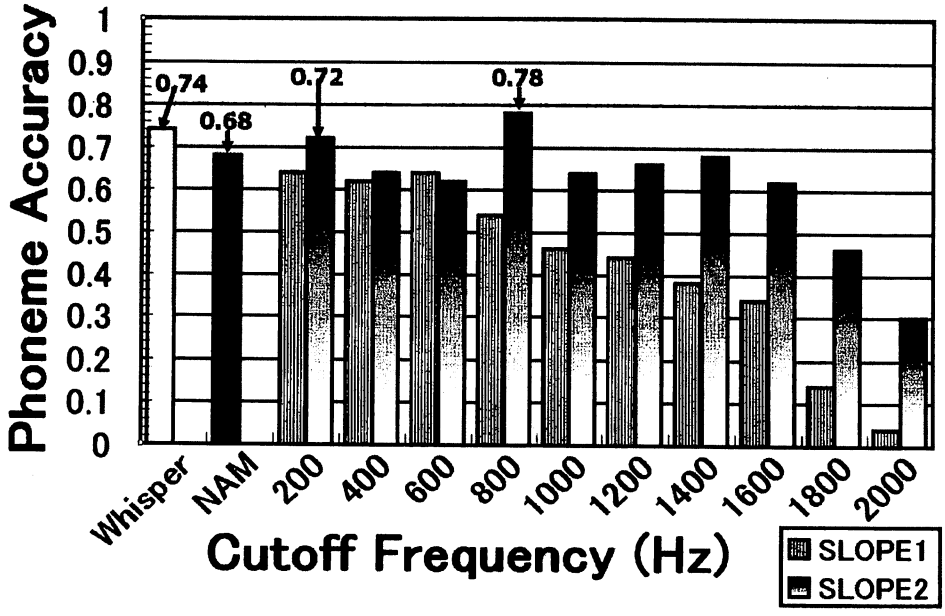

Figure 6: 人間の聞き取り認識精度（無意味単語）

7. まとめ

以上の実験ではソフトウェア的なデジタルハイパスフィルタを用いたが，デモンストレーションなどでは自作NAMマイクアンプ出力にCR回路（微分回路）でアナログハイパスフィルタ（1kHz, 6dB/oct）を組み込んでHPF-NAMを出力している．わずかに抵抗とコンデンサ一個ずつで非常に低コストの疑似ささやき声化が可能であり，学習もいらず，ディレイもない．周囲に聞き取られない無音声通信のデモンストレーションなどで出力をリアルタイムで聴取してもらっても，オリジナルNAMを聴取してもらっていた時は，挨拶や簡単な質問や命令など電話会話として普遍的で予想の付く発話内容を伝えていたが，HPF-NAMを使い出してから一般的な会話や電話番号など数字の聞き取りも可能となった．

今回の男性話者一名のNAM発話による実験では，自然性と聞き取り認識精度の両方を勘案して，SLOPE2タイプのなだらかなスロープ特性で，800～1200Hzのカットオフ周波数が，聞き取りやすい疑似ささやき声化に適していると思われる．

現在，機械認識での認識率に変化が出るかどうか，実験中である．今後，NAM話者の性別や年齢や言語による違いがあるかどうか，声質変換によりささやき声化したサンプルとの音質比較などを検討する必要がある．

参考文献

[1] 中島淑貴，柏岡秀紀，ニックキャンベル，鹿野清宏 “非可聴つぶやき認識”，信学会誌, 87(9) 1757-1764, 2004.

[2] 中島淑貴，柏岡秀紀，ニックキャンベル，鹿野清宏 "非可聴つぶやきをインターフェースとするコミュニケーションのためのソフトシリコーン型NAMマイクロホンの開発", 信学会誌, 89(8) 1757-1764, 2006.

[3] T. Toda and K. Shikano, “NAM-to-Speech Conversion with Gaussian Mixture Models”, Proc Eurospeech, 2005.

[4] 中桐，戸田他，“NAMマイクを用いて収録した無声音声の品質改善”，日本音響学会2006年春季研究発表会講演論文集, 1-4-15, 2006.

（本研究は平成17年度総務省SCOPE-S『発声障害者の音声コミュニケーション手段の研究開発』により実施した）