サウンドマスキングシステムにおける データベースを用いたマスカー作成法の提案*

☆三戸武大，荒井隆行，安啓一（上智大・理工）

1　はじめに

スピーチプライバシーの保護を実現する方法の一つとして，サウンドマスキングがある[1]。マスキングの手法として，従来は，定常雑音を使って音声をマスクすることが多かった[2]。しかし最近では，定常雑音を使用するよりも，話者の音声を加工したマスカーを使用する方が，マスキング効率が高まるという報告がなされている[3,4]。話者の声を使用したマスカー作成法の一つに，マイクロホンから取り込んだターゲット話者の音声を実時間処理により加工するものがある。例えば中嶋ら[5]は，実時間処理を考慮に入れた上で，Ito *et al*. [3]や Arai [6]によって提案されたマスカーの評価を行った。また赤木ら[7]は，実時間処理により，ターゲット音声の基本周波数を一致させ，スペクトル包絡を変形したマスカーを提案した。上記 2 つの報告では，発話者の声が加工されマスカーとして流されることから，高いマスキング効率が得られる[3,4]。しかし，自分自身の声がマスカーとして使われることに抵抗を示したり，たとえ加工されていたとしても，聞かれたくないターゲット音声そのものをマスカーとして流すことに抵抗を持ったりする利用者もいる可能性がある。

そこで本報告では，サウンドマスキングシステムにおいて，様々な音が収められたデータベースを予め用意し，ある基準に従って音をデータベースから選択することでマスカーを作成する方法を提案する。その際，ターゲット音声の基本周波数とスペクトル形状がどのように結果に影響するかを調べるため，複数のマスカーによる単語了解度試験を行い，各マスカーの評価を行った。

2　本報告で使用したマスカー

本報告では，6 種類のマスカーを使用した。以下にマスカーの詳細について述べる。本研究では最終的に実時間処理を視野に入れており，実際には以下に示す手順で処理を施すことを想定している。しかし，本報告における実験では，簡単のためオフラインで処理を行った。

I.　マスカーとして使用したい音（以下，エントリー）と共に，必要に応じてエントリー毎の基本周波数の平均値，ケプストラム等の特徴量を，データベース内に用意する。

II.　実時間処理によりマイクロホンから取り込んだターゲットをフレーム毎に分析し，上記の特徴量を得る。

III.　ターゲットの分析結果から，基準を最も満たしたエントリーを特定のアルゴリズムによりデータベースから選択し，これをマスカーとする。

なお，IIIの基準については大きく分けて 4 種類を試した。以下に，それらの基準によって得られる 4 種類のマスカーについて述べる。

2.1 F0 マスカー

赤木ら[7] を参考に，F0 マスカーを作成した。以下に作成手順を示す。なお，①，②の処理は予め行い，データベース内にデータを格納した。その上で，③以降の手順に従いマスカーを作成した。

①　親密度別単語了解度試験用音声データセット（FW03）[8]の音声から，女性話者 fto の単音節音声を 100 個選択し，同じ時間幅のフレームに分割した。フレーム長は可変ではあるが，本報告では 100 ms とした。なお，本報告ではターゲットを女性話者の音声に限定したので，データベースも女性話者の音声のみとした。さらに音響分析ソフトウェア Praat [9]を使用して，それぞれのフレームの基本周波数の平均値を操作（原音を±25 Hz, ±50 Hz, ±75 Hz）したものを新たに作成し，それらを加えたすべてのエントリーをデータベース内に用意した。データベースの全

* Creating masker using a database in a sound masking system, by SANNOHE, Takehiro, ARAI, Takayuki, and YASU, Keiichi (Sophia University).

エントリー数は約 2400 個であった。

② データベース内の各エントリーの基本周波数の平均値を Praat により計算し，データベース内にその値を格納した。ただし，無声音に関しては，単音節/sa/の摩擦部分/s/を無声音用のエントリーとすることで対応した。

③ ターゲットに関して，フレーム毎に基本周波数の平均値を計算した。本報告では予め Praat によって計算しておいた。

④ ターゲットに対して，フレーム毎にデータベース内のすべてのエントリーと基本周波数の平均値の差を計算した。最も差が小さかったエントリーをそのフレームに対するマスカーとして選択した。無声音に関しては，無声音用のエントリーを選択した。

⑤ フレーム毎に④の作業を繰り返し，選択されたエントリーを順次連結したものを信号Aとした。選択されたエントリーを順次連結する際，李ら[10]を参考にし，ターゲットのレベルにマスカーのレベルを追従させた。実際には，ターゲットの各フレームと対応するエントリーの実効値が等しくなるようにレベルを合わせた。

⑥ 信号Aとは別に，1/2 フレーム遅らせた時点から③～⑤と同様の処理を行った音声を作成した。これを信号Bとした。

⑦ 信号 A と信号 B を加算し， F0 マスカーとした。

2.2 SPECマスカー

スペクトル包絡をターゲットと類似させることでSPECマスカーを作成した。マスカーの作成手順は 2.1 節の手順と同様にしたが，②～④を以下の②'～④'に置き換えた。

②' データベース内の各エントリーの FFT ケプストラムの低ケフレンシ部（1 次～30 次の項）を MATLAB により計算し，データベース内にその値を格納した。

③' ターゲットに関して，フレーム毎に FFT ケプストラムの低ケフレンシ部を計算した。本報告では予め MATLAB によって計算しておいた。

④' ターゲットに対して，フレーム毎にデータベース内のすべてのエントリーと FFT ケプストラムの低ケフレンシ部の 2 乗誤差を計算した。最も差が小さかったエントリーをそのフレームにおけるマスカーとして選択した。

2.3 F0_SPECマスカー

基本周波数とスペクトル包絡をどちらも考慮することで$F0_SPEC_n$（$F0\_SPEC_{near}$）マスカー，$F0\_SPEC_f$（$F0\_SPEC_{far}$）マスカー，$F0\_SPEC_m$（$F0\_SPEC_{middle}$）マスカーを作成した。上記 3 種のマスカーは，基本周波数が近接する候補の中でスペクトル包絡の距離がマスキング効率に与える影響を調査することを目的とした。マスカーの作成手順は，基本的には 2.1 節の手順と同様にしたが，②，③におけるエントリーとターゲットの基本周波数の平均値の計算に加え，FFTケプストラムの低ケフレンシ部も計算した。また，④を次のようにした。3 種のF0_SPECマスカーいずれにおいても，まずターゲットの注目するフレームにおける基本周波数の平均値が近接しているエントリーを候補として複数選択した。その際，赤木ら[7]の知覚的融合が成り立つ条件を参考に，±2 Hzのエントリーすべてを候補とした。選択されたエントリーにおいて，さらにターゲットの当該フレームとのスペクトル距離が最も近いエントリーを$F0_SPEC_n$マスカーとした。また，ターゲットの当該フレームとのスペクトル距離が最も遠いエントリーを$F0_SPEC_f$マスカーとした。さらに，選択されたすべての候補とターゲットの当該フレームとのスペクトル距離の平均値を求め，その平均値と最も近いものを$F0_SPEC_m$マスカーとした。

2.4 RANDOMマスカー

Ito *et al*. [3] を参考に，基本周波数もスペクトル包絡も考慮せずにデータベースの内のエントリーを無作為に選択したものをRANDOMマスカーとした。作成手順は2.1節の手順と同様にしたが，②，③でエントリーとターゲットの特徴量を計算せず，④のエントリーの選択を無作為にすることでマスカーを作成した。

3 実験

3.1 実験参加者

日本語を母語とする健聴者，男性 11 名，女性 7 名，計 18 名（平均 21.6 歳）が実験に参加した。健聴か否かは，参加者の自己申告で確認した。

3.2 刺激音

ターゲットには，親密度別単語了解度試験用音声データセット（FW03）[8]から，女性話者 fhi の音声を採用し，その中の親密度 7.0~5.5 の単語群から 240 語を選択した。これらはすべて 4 モーラからなる単語である。上記のコーパスの音声はサンプリング周波数 48 kHz，量子化ビット数 16 bit のものであるが，本報告ではサンプリング周波数を 16 kHz にダウンサンプリングし使用した。また FW03 は 1 セット 50 単語で音素バランスが取れている。そこで本報告では，6 種類のマスカーそれぞれにつき 8 単語を用い，50 単語中の 48 単語を使用することで，なるべく音素バランスを崩さないように配慮した。それらをターゲット対マスカー比（target-to-masker ratio，以下 TMR）条件毎に用意し，全体で 240 単語（8 単語×マスカー6 種類× TMR 5 条件）を用いた。

ターゲットの呈示レベルは，実験参加者の頭部中央を想定した位置（高さ 1.1 m）で騒音レベル（A 特性）が 50 dB となるように騒音計（リオン・精密騒音計 NL-32）で設定した。マスカー作成時のフレーム長は，全ターゲットの平均音素長を参考に 100 ms とした。各マスカーの呈示レベルは，TMR が－15 dB, －10 dB, －5 dB, 0 dB, 5 dB になるように調整した。全参加者におけるターゲットとマスカーの組み合わせについては，TMR 毎に用意した 48 単語（8 単語×マスカー6 種類）を 8 単語ずつの 6 グループに分け，グループ毎にマスカーを割り当てた。各単語がすべてのマスカー条件において割り当てられるよう，参加者が変わる毎に，割当てを変えた。マスカーが 6 種類なので，参加者 6 人でマスカーの割当てが 1 周するようにし，TMR 毎にカウンターバランスを取った。参加者は 18 人いたので，マスカーの割当ては 3 周した。

3.3 実験手順

実験は防音室で行われた。Fig. 1 に実験環境を示すとともに，以下に実験の詳細について述べる。実験参加者は，高さ 1.0 m の台に設置されたスピーカ（ヤマハ・MSP-3）から 1.8 m 離れた位置に着席し，スピーカから呈示される刺激音を聴取した。刺激音は，ターゲットとマスカーを予めMATLABを用いて加算したモノラル音をスピーカから呈示するものとし，各刺激音の呈示は 1 回のみとした。一つの刺激音の呈示が終了する毎に，目の前に設置されたノートPCに聴取した音声をひらがなでタイプ入力してもらった。30 条件（マスカー6 種類×TMR 5 条件）に対し，条件毎に 8 単語を使用したため，実験参加者は，全体で 240 刺激を聴取した。刺激の呈示順はTMR毎にランダムに並べ替えた。実験は，TMRの小さい方から，－15 dB, －10 dB, －5 dB, 0 dB, 5 dBの順番で行った。

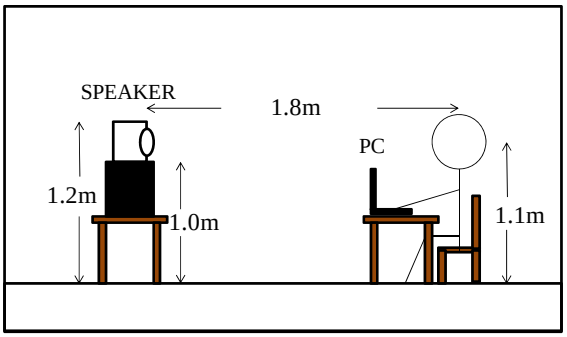

Fig. 1 : Experimental setup

3.4 評価基準

実際に呈示した刺激と，参加者がタイピングしたひらがなが完全に一致した場合を正解とした。ただし，文字そのものは違うが，現代の日本語では音声が一致する場合（例えば，「きずぐち」が正解で「きづぐち」と回答した場合等）は正解とした。そしてマスカー6 種類×TMR 5 条件の計 30 条件の各々に関して，実験参加者 18 名が回答した 8 単語の正答率の平均を単語了解度（単位は%）として評価した。なお，$F0\_SPEC_n$ マスカー，$F0\_SPEC_f$ マスカー，$F0\_SPEC_m$ マスカーに関しては，実験後に不具合を確認したため評価の対象外として対応した。

4　実験結果及び考察

Fig. 2に，実験の結果を示す。それぞれの図中の曲線は，マスカーの呈示レベルと単語了解度との関係をロジスティック関数による回帰分析によって求めたものである。

Fig. 2によると，基本周波数の平均値を類似させたF0 マスカーの曲線は，無作為に作成したRANDOMマスカーの曲線よりも下方に位置している。したがって，F0 マスカーはRANDOMマスカーに比べて，マスキング効率が高いことが確認できる。マスキング効率を高めるには基本周波数をターゲットに類似させることが有効だと考えられる。

一方，スペクトル包絡を類似させた SPEC マスカーの曲線は，無作為に作成した RANDOM マスカーの曲線よりも上方に位置している。したがって，SPEC マスカーのマスキング効率は必ずしも高くないことが確認できる。理由として，ターゲットとマスカーのスペクトル形状を類似させた結果として音自体が近似しすぎてしまい，ターゲットをターゲットでマスクしようとするような状況が生じていたことが考えられる。

本報告では， F0_SPEC マスカーに関して実験後に不具合を確認したため，結果を評価することができなかった。しかし仮説として，スペクトル距離を考慮した F0_SPEC マスカーの方が，考慮していない F0 マスカーよりも，マスキング効率が高くなると考えている。F0_SPEC マスカーに関しては再実験を行い，マスキング効率を高めるにはスペクトル形状がターゲットにどのように類似していることが有効なのか調査したいと考えている。

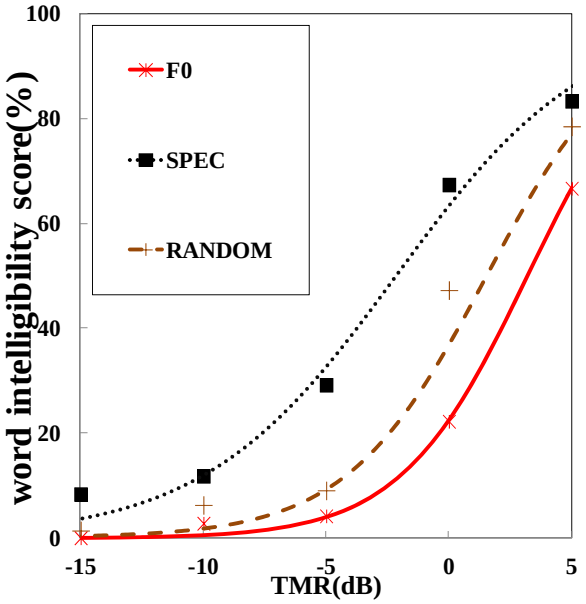

Fig. 2 : Result of the word intelligibility test

5　おわりに

本報告では，サウンドマスキングにおいて，様々な音が収められたデータベースを用意し，ある基準に従って音をデータベースから選択することでマスカーを作成する方法を提案した。本報告の提案法を用いて作成した 6 種類のマスカーによる単語了解度試験を行い，そのうちの 3 種類のマスカーについてマスキング効率の評価を行った結果，マスキング効率を高めるには，基本周波数を類似させることが有効であるという結果を得た。

今後の課題として，本報告で提案したアルゴリズムを用いて作成したマスカーを DSP に実装し，評価実験を行うことが挙げられる。実時間処理に向けて，DSP 等に実装することを考える上で，データベースの規模が重要になってくる。データベースの規模を大きくする程，DSP の計算量が多くなり，時間遅延が大きくなってしまう。赤木ら[7]の報告では，時間遅延を 10 ms 以内にすることが望ましいとされている。なるべく時間遅延を小さくするために，エントリーの厳選をすることが必要だろう。また，エントリーの内容について，本報告では女性の音声しか用意しなかったが，複数の人間の音声はもちろん，環境音や定常雑音等も加えて用意することでマスキング効率が高まる可能性もある。さらに時間反転処理を含むその他の処理を施したエントリーもデータベースに含めることも検討しているが，データベースの内容に関しては今後検討の余地があると考えられる。

最後に，本報告ではマスキング効率の評価のみでアノイアンスに関しての評価を行っていない。今後，本報告のマスカーを使用したサウンドマスキングシステムの実用化に向け，アノイアンスの評価実験を行う予定である。

謝辞

本研究の一部は文部科学省私立大学学術研究高度化推進事業上智大学オープン・リサーチセンター「人間情報科学研究プロジェクト」の支援を受けて行われた。

参考文献

[1] 佐藤, 清水, 音響誌, 64(8), 475-480, 2008.
[2] 佐伯他，信学技報，EA 103 (398)，43-48, 2003.
[3] Ito *et al.*, Proc. INTER-NOISE, 2007.
[4] Ueno *et al.*, Proc. INTER-NOISE, 2007.
[5] 中嶋他，音講論（秋），1145-1148, 2009.
[6] Arai *et al.*, Acoust. Sci. Tech, 31 (2), 188-190, 2010.
[7] 赤木，入江， 信学技報，EA2011-99, EMM2011-59, 2011.
[8] 天野他，NII 音声資源コンソーシアム, 2006.
[9] Boersma & Weenink, Glot International, 5 (9), 341-345.
[10] 李他，音講論（秋），1077-1078，2010.