混合音声からの特定話者信号抽出に関する基礎的考察

薄 田 憲 宏[*1]・大 島 静 夫

A Basic Study on the Extraction of a Specific Speaker's Signals from the Male-Female Mixed Speech

Norihiro USUDA* and Shizuo OHSHIMA

(2002年11月30日受理)

Recently, many researches have been reported on the separation and extraction method to abstract a specific speaker's speech at the conference. This study focuses on the basic method to separate and extract a specific speaker's speech on the premise that there are independent speech signals of male and female speakers in the signals, with a clue of the characteristic for the speech signal in a frequency domain. On a concrete basis, it is the method to extract a specific speakers syllable by (1) discerning the monosyllable that constitutes the objective syllable for extraction from other independent speech signals, (2) defining the average value of its Short Time Fourier Transform as the specific filter of the speaker for the monosyllable and (3) filtering the STFT spectra of the male-female mixed speech through the above-mentioned specific filter of the speaker. The result of a numeric experiment shows that this method can extract the speech in question despite remaining the speech of speakers other than the objective speaker for extraction to an ignorable extent.

1. はじめに

複数の話者が同時に発言する会議のような場において，特定の話者の音声を取り出す音声分離抽出手法に関し，近年多数の報告がなされている [1] [2]。[1] では 2 つの音源からの信号を 2 個のマイクロホンで受音し，各受音信号の短区間フーリエ変換（以降 STFT）スペクトルの比を求め，その値より各フレームの周波数成分がどちらの音源から発せられたのかを判定し，分離している。また [2] では，男女混合音声を一点受音し，STFT を行うデータ個数が異なるウィンドウを用いる Multi windowed STFT によって受音信号の基本周波数を求め，男女では基本周波数が違うことを利用して音声分離を行っている。

本論文では，1 チャンネルの信号中に男女各話者の独立した信号があることを前提にし，その音声の周波数領域における特徴を手がかりとし，男女混合音声から特定話者の音声を分離抽出する基本的手法を検討した。

具体的には，抽出対象音節が単音節に分解可能であるとし，抽出対象の音節に含まれる単音節と同じ音の単音節を他の音節から切り出し，それらの STFT スペクトルの平均値を音声に対するフイルタとし，男女混合音声の STFT スペクトルに掛け合わせることによって特定話者の音声信号を抽出する手法である。

2. 音声フィルタを用いる抽出法

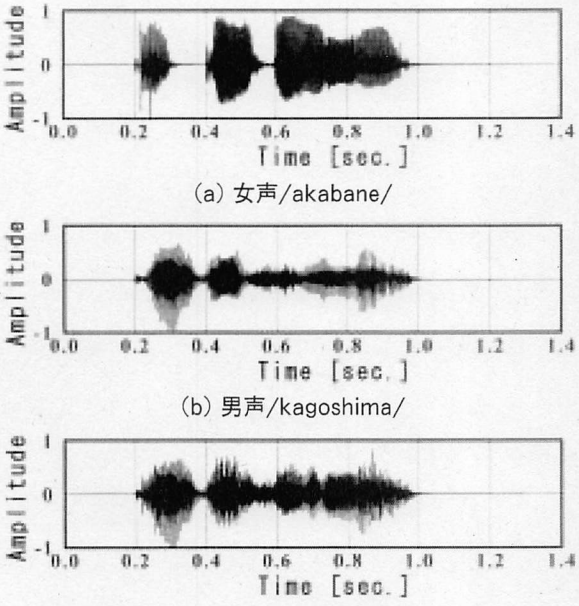
2.1 解析に用いた音声

本研究では，解析音声としてアプリケーションソフト Acoustic Core Version 3.0（Arcadia社）に含まれる地名音声ファイル（総音節数2550）から任意に抽出した音節を使用した。使用音節のサンプリング周波数は16000Hz，量子化数は16bit である。各音節は，データ後部に無音部分を付け加えることでデータ長を Ndata＝20936に調整している。

図 1 に以降の解析で用いる代表的な音声信号として，(a) 女声/akabane/，(b) に男声/kagoshima/，(c) に (a) と (b) の男女混合音声信号の時間波形を示す。

---

* 秋田高専専攻科学生



(a) 女声/akabane/

(b) 男声/kagoshima/

(c) (a)(b)の男女混合音声

図1　代表的な音声信号の時間波形

2.2　解析に用いた記号

解析音声信号は離散化されているため，その時系列順にデータ番号を i として $v[i]$ と表す。また v_m のように，添字 m は男声，f は女声，b は男女混合音声信号を表すことにする。

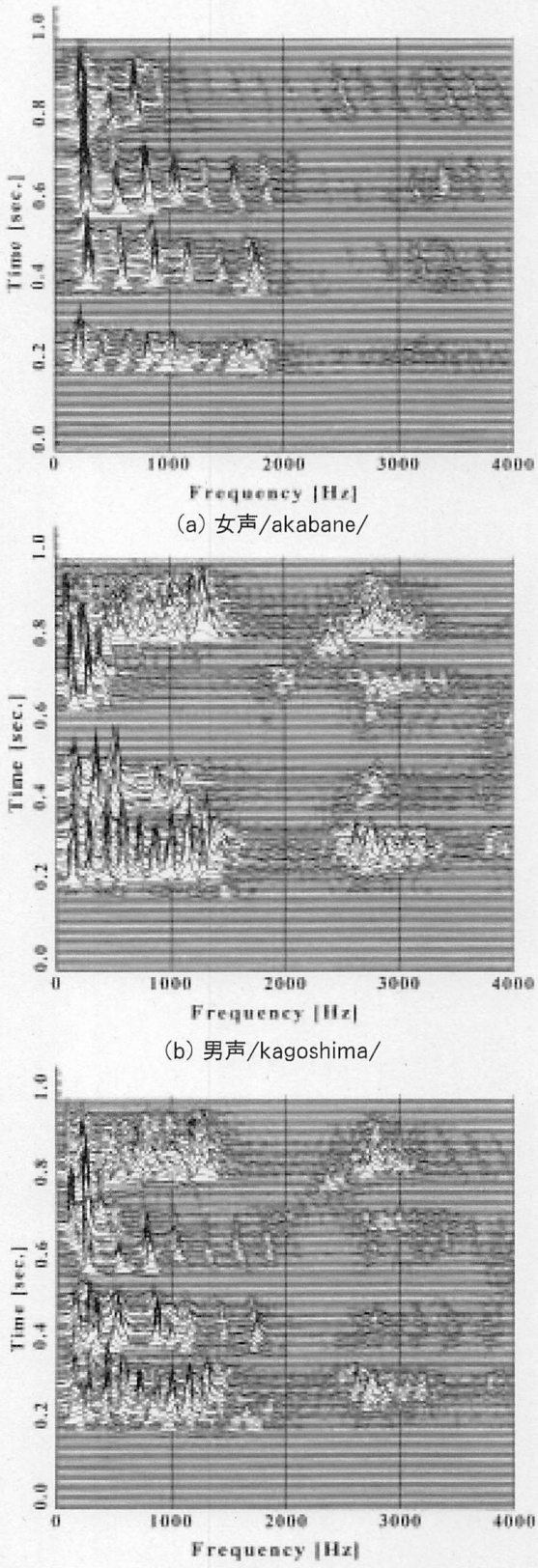
解析にはオーバーラップを伴うSTFTを用いた。STFT解析に用いるデータ数 N は512，オーバーラップ数は384，総フレーム数 M は156とした。表1に，音声信号 $v[i]$ とSTFT解析のための配列 $V[f_r,n]$ の関係を示す。

表1　STFT解析のための音声信号配置表

フレーム番号 f_r	フレーム内データ番号 1	2	…	512
1	$V[1,1]=v[1]$,	$V[1,2]=v[2]$,	…	$V[1,512]=v[512]$
2	$V[2,1]=v[129]$,	$V[2,2]=v[130]$,	…	$V[2,512]=v[640]$
:	:	:	:	:
f_r	$V[f_r,1]=v[19841]$,	$V[f_r,2]=v[19842]$,	…	$V[f_r,512]=v[20532]$

ここで，音声信号 $V[f_r,n]$ のSTFTを $S[f_r,k]$ とすると，各フレームのSTFTは(1)式で与えられる。kは直流を含めた高調波成分の次数を示す記号とする。

$$S[f_r,k]=\sum_{n=1}^{N} V[f_r,n]\exp\{-\frac{j2\pi}{N}(k\text{-}1)(n\text{-}1)\} \quad \cdots\cdots\cdots(1)$$
$$k=1,2,3\cdots N$$



(a) 女声/akabane/

(b) 男声/kagoshima/

(c) (a)(b)の男女混合音声

図2　代表的な音声信号のSTFTスペクトル

図1で示した音声のSTFTスペクトルを図2に示す。

また，STFTスペクトル$S[f_r,k]$の逆STFT変換（以降ISTFT）は，(2)式で与えられる。

$$V[f_r,n]=\frac{1}{N}\sum_{k=1}^{N}S[f_r,k]\exp\{\frac{j2\pi}{N}(k\text{-}1)(n\text{-}1)\} \quad \cdots\cdots(2)$$
$$n=1,2,3\cdots N$$

今，男女混合音声信号が，任意のフレームにおいて，(3)式で与えられものとすると，そのSTFTは(4)式となる。

$$V_b[f_r,n]=V_m[f_r,n]+V_f[f_r,n] \quad \cdots\cdots(3)$$
$$S_b[f_r,k]=S_m[f_r,k]+S_f[f_r,k] \quad \cdots\cdots(4)$$

(4)式より，女声信号を抽出するときは，$S_f[f_r,k]=S_b[f_r,k]\text{-}S_m[f_r,k]$の計算を行い，ISTFTを施し時間波形に戻すと，100%完全に目的の音声信号を抽出できる。しかしながら，この$S_m[f_r,k]$を推定するためには，高度な計算が要求される。この$\hat{S}_f[f_r,k]=System\{S_b[f_r,k]\}$（以降$\hat{}$は推定値）を満足する簡単な*System*を音節フィルタを用いて作成することが本論文の目的である。

2.3 同一音節フィルタを用いる抽出法

最終的に検討したい手法は，抽出対象信号に含まれる単音節と同じ音としての単音節を他の音節から切り出し，そのSTFTスペクトルを音声抽出フィルタとして用いる手法である。この手法の妥当性の確認のため，ここでは同一音節を音声フィルタとして用いる場合について検討する。以降このフィルタを同一音節フィルタと呼ぶ。

同一音節フィルタ使用時のスペクトル抽出手順を以下に示す。またフレーム番号52を例とし，各手順におけるスペクトルの推定過程を図2に示す。

手順1：男女混合音声信号をSTFTする。図3(a)。
手順2：抽出対象話者の同一音節フィルタを作成する。

同一音節フィルタは，抽出対象話者のSTFTスペクトルを最大値で正規化したものであり，女声信号を抽出する場合(5)式で与えられる。図3(b)。

$$W_f[f_r,k]=|S_f[f_r,k]|/\max(|S_f[f_r,k]|) \quad \cdots\cdots(5)$$

手順3：排除対象話者同一音節逆フィルタを作成する。

今の場合排除対象である男声信号を正規化する。$[f_r,k]$個で構成されるの1のマトリックスを$1[f_r,k]$

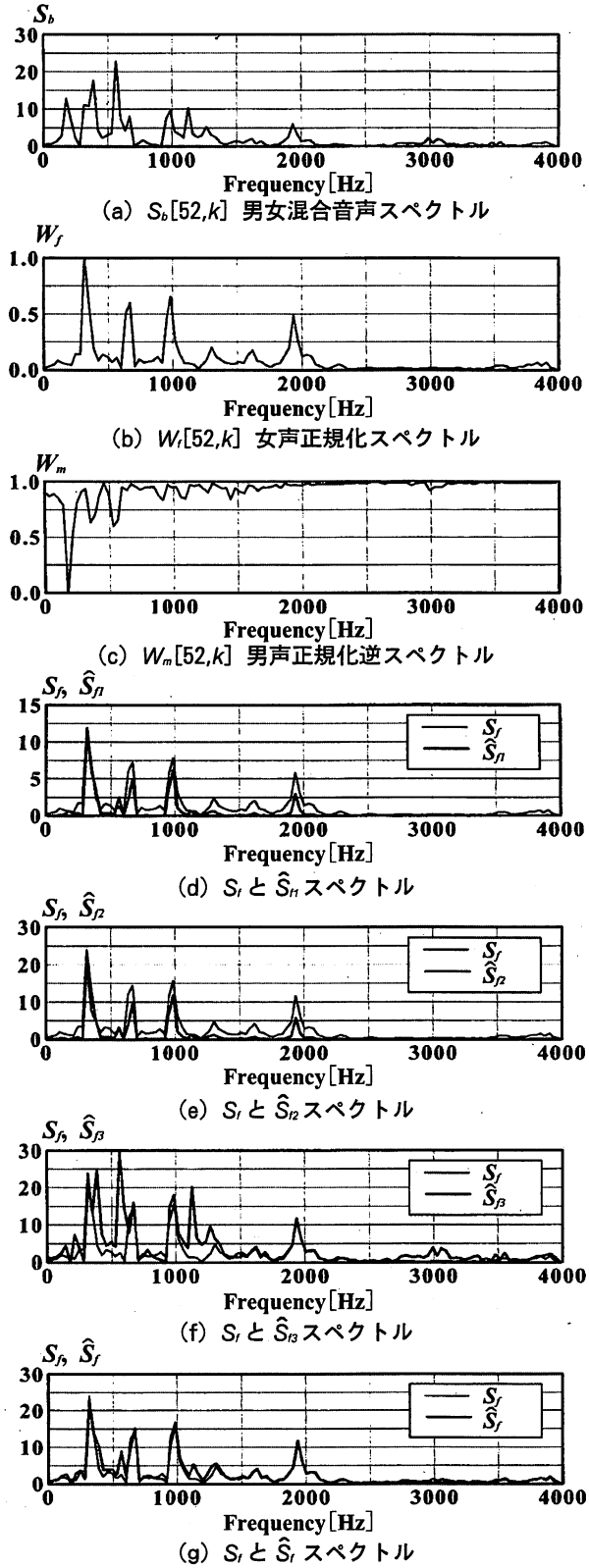


図3　手順とスペクトルの関係

と表現すると，逆フィルタは，(6)式で与えられる。図3(c)。

$$W_m[f_r,k]=1[f_r,k]-|S_m[f_r,k]|/\max(|S_m[f_r,k]|) \quad \cdots\cdots(6)$$

手順4：男女混合音声信号のSTFTスペクトルに同一音節フィルタを掛け合わせたスペクトルを $\hat{S}_{f1}$ とおく。この操作は，抽出したい音節を強調的に取り出すことを意味する。図3(d)。

$$\hat{S}_{f1}[f_r,k]=S_b[f_r,k]\cdot W_f[f_r,k] \quad \cdots\cdots(7)$$

さらに，$\hat{S}_{f1}$ に手順3の男声同一音節逆フィルタを掛け合わせたスペクトルを $\hat{S}_{f2}$ とおく。この操作は，男声を強調的に打ち消すことを意味する。図3(e)。

$$\hat{S}_{f2}[f_r,k]=\hat{S}_{f1}[f_r,k]\cdot W_m[f_r,k] \quad \cdots\cdots(8)$$

信号排除機能の確認のため，図3(f)に男女混合音声信号に $W_m[f_r,k]$ のみをかけたスペクトル $\hat{S}_{f3}$ を示す。

$$\hat{S}_{f3}[f_r,k]=S_b[f_r,k]\cdot W_m[f_r,k] \quad \cdots\cdots(9)$$

手順5：手順2および3の改良として，同一音節フィルタに閾値 Th を定め，レベルの小さいスペクトルを0とする。女声音声信号の場合のアルゴリズムを以下に示す。

$$W_f[f_r,k]=\begin{cases} if\ W_f[f_r,k]<Th,\ W_f[f_r,k]=0 \\ otherwise,\ W_f[f_r,k]=W_f[f_r,k] \end{cases}$$

手順6：(8)式に(5)(6)(7)式を代入し，$[f_r,k]$ を省略した形式で式を表現し直すと(10)式となる。

$$\hat{S}_{f2}=(|S_m+S_f|)\frac{|S_f|}{\max(|S_f|)}\cdot\left(1-\frac{|S_m|}{\max(|S_m|)}\right) \quad \cdots\cdots(10)$$

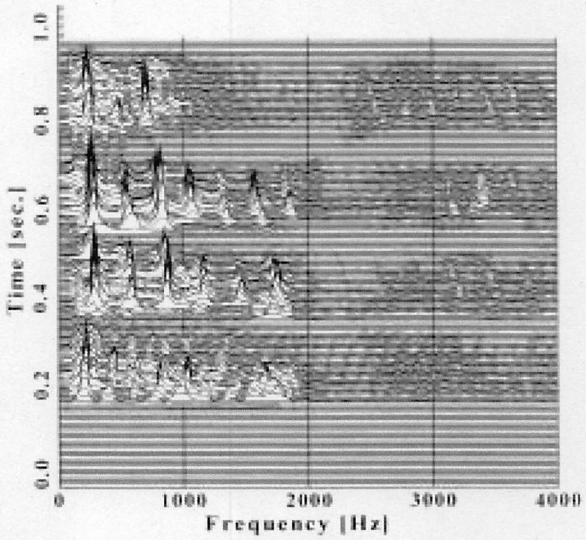
ここで S_f と S_m は，周波数領域でレベルのピークの大部分が一致しないとすると，$S_f \gg S_m$ なら $S_m \fallingdotseq 0$，$S_m \gg S_f$ なら $S_f \fallingdotseq 0$ となり，$S_f \cdot S_m \fallingdotseq 0$ と仮定できる。この性質を利用すると，

$$\hat{S}_{f2} \fallingdotseq (|S_m+S_f|)\frac{|S_f|}{\max(|S_f|)} \fallingdotseq \frac{|S_f|^2}{\max(|S_f|)} \fallingdotseq \frac{|S_f|^2}{\max(|S_b|)}$$

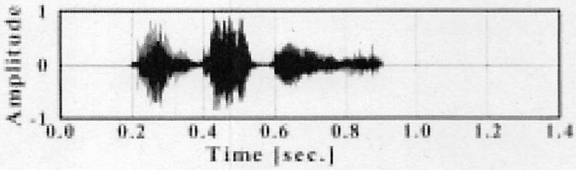
これより S_f は，(11)式のように推定できる。(11)式による推定値を図3(g)に示す。

$$|\hat{S}_f|=\sqrt{\hat{S}_{f2}\cdot\max(|S_f|)} \quad \cdots\cdots(11)$$

手順7：同一音節フィルタを掛け合わせたスペクトルをISTFTし，時間波形に戻す。



(a) 女声/akabane/の抽出STFTスペクトル



(b) 女声/akabane/の抽出時間波形

図4　同一音節フィルタによる抽出

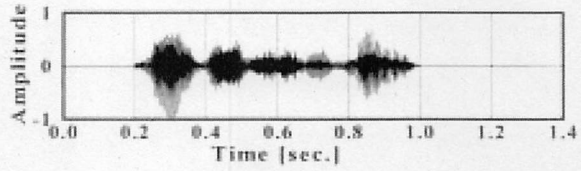


図5　男声/kagoshma/の抽出時間波形

$$\hat{V}_f[f_r,n]=\mathrm{ISTFT}(\hat{S}_f[f_r,k]) \quad \cdots\cdots(12)$$

図4(a)に，本節の手法により抽出した女声/akabane/のSTFTスペクトルを，図4(b)に時間波形を示す。

図2(a)(b)に示したSTFTスペクトルと比較すると，1000Hz以下の女声信号が抽出されており，また2000Hz以上の男声信号は排除できていることがわかる。図1(b)に示した時間波形と比較すると，波形の概形はほぼ一致し，良好に信号が抽出できているといえる。

同様に図5には，同様に男声/kagoshima/の抽出時間波形を示す。

2.4　同一単音節平均フィルタを用いる抽出法

前節より，同一音節フィルタを用いることで音声

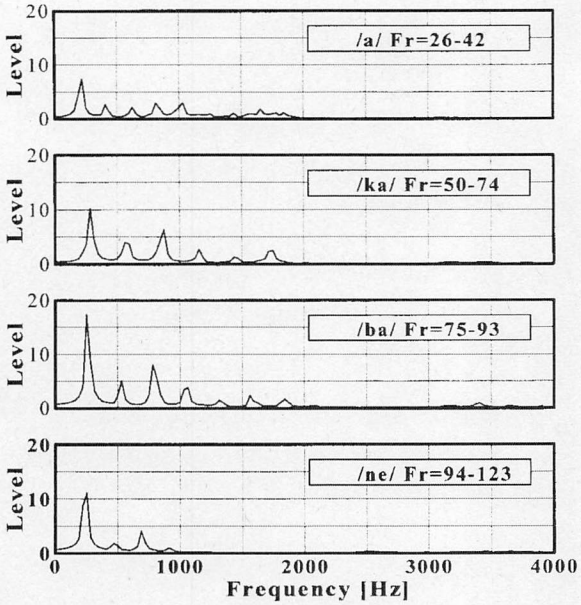


図6　同一単音節平均フィルタによる抽出

信号抽出が可能であることが確認された。これを他音節から切り出した単音節より作成したフィルタにより音声抽出を行うのが最終目標である。そこで本節では，前節の手順2で作成したフィルタを単音節ごとに平均したフィルタに置き換えることによっても，音声信号抽出が可能かを検討した。本節では，手順2は次のように書き改められる。手順3も同様である。

手順2：抽出対象話者の単音節ごとに平均を取った単音節平均フィルタを作成する。計算式を(13)に示す。

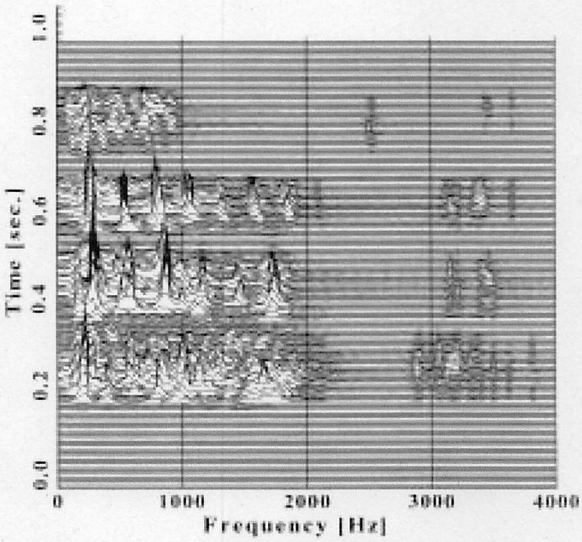
$$W_f[f_n,k]=\sum_{f_r=a}^{b}|S_f[f_n,k]|/\max\left(\sum_{f_r=a}^{b}|S_f[f_n,k]|\right) \quad \cdots\cdots(13)$$

ここで a は単音節が属する先頭フレーム

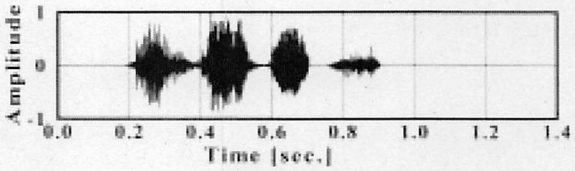
b は単音節が属する最終フレーム　とする。

以降このフィルタを同一単音節平均フィルタと呼ぶ。図6に，各単音節毎の平均スペクトルを示す。

図7に，本節の手法により抽出した女声/akabane/のSTFTスペクトルと時間波形を示す。図2のスペクトルと比較すると，2000Hz以上の男声信号を排除しているが，1000Hz以下の女声信号に多少男声信号が混じっていることがわかる。しかし，全体として女性の持つ周期的ピーク構造は壊れておらず，聴覚上では，男声が多少残り音質も悪くはなるが，女声音声としてはっきりと認識できる。従って，同一単音節平均フィルタでも，音声分離抽出が可能であるといえる。図8に同様に男声/kagoshima/の抽出音声信号の時間波形を示す。



(a) 女声/akabane/の抽出 STFT スペクトル



(b) 女声/akabane/の抽出時間波形

図7　同一単音節フィルタによる抽出

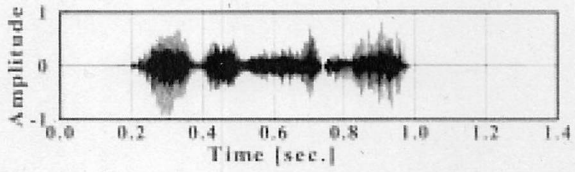


図8　男声/kagoshima/の抽出時間波形

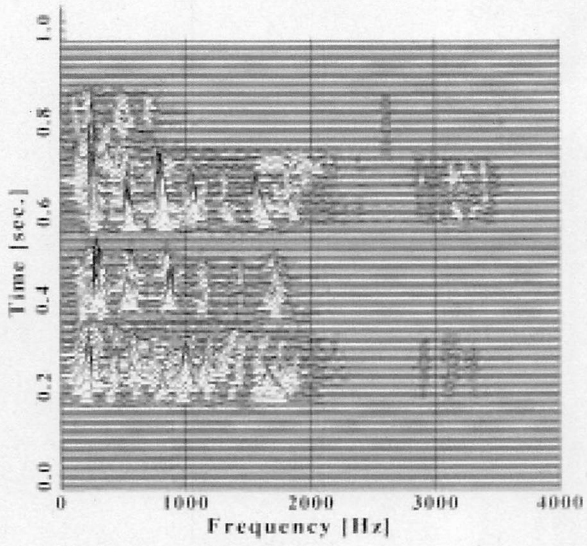
2.5　他単音節平均フィルタを用いる抽出法

前節より，同一単音節平均フィルタを用いても音声信号抽出が可能であることが確認された。最後に一般的な手法として，他音節から切り出した単音節により作成したフィルタにより，音声信号抽出を行う方法について述べる。以降このフィルタを他単音節平均フィルタと呼ぶ。

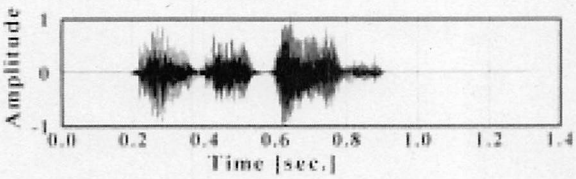
手順2は以下のように変更される。

手順2：他音節より，抽出対象音節に含まれる単音節を切り出す。切り出した音節をSTFTし，各フレームごとの周波数成分の平均値を求め正規化したスペクトルの他音節平均フィルタを作成する。

図9に，他単音節平均フィルタにより抽出した女声/akabane/のSTFTスペクトルと時間波形とを



(a) 女声/akabane/の抽出 STFT スペクトル



(b) 女声/akabane/の抽出時間波形

図9　他単音節フィルタによる抽出

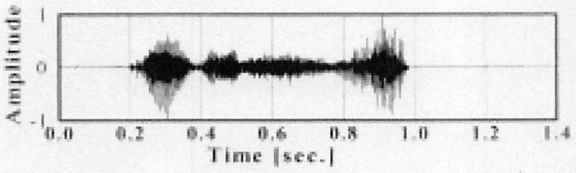


図10　男声/kagoshima/の抽出時間波形

示す。ここで使用した他音節は女声音声信号の，/akabaneseN/の/a/と/ka/，/baNkoku/の/ba/，/neyagawa/の/ne/である。同様に男声信号においては，/kainaN/の/ka/，/koNgo/の/go/，/ooshima/の/shi/，/ooshima/の/ma/である。本手法においても，抽出したSTFTスペクトルにおいて女声の周期的ピーク構造は保たれている。しかし，男声のものと思われるスペクトルは完全には排除できず，このために抽出音節には男声の影響が多少残っている。

図10には男声/kagoshima/の抽出音声信号の時間波形を示す。

3. 結論

すでに録音されている記録媒体中の男女混合音声信号から一方を分離抽出することを目的に，音声フィルタを作成し，それらを男女混合音声信号のSTFTスペクトルに掛け合わせることによって音声分離抽出を行った。

ここで用いた最終手法は

①男女混合音声信号をSTFTする。

②他音節より，抽出対象音節に含まれる単音節を切り出す。これらの単音節のSTFTスペクトルの平均値を求め，他単音節平均フィルタを作成する。

③同様に，排除対象話者の他単音節平均逆フィルタを作成する。

④男女混合音声信号のSTFTスペクトルに②③の他単音節平均フィルタを掛け合わせる。

⑤推定スペクトルの平方根をとる。

⑥推定スペクトルをISTFTする。

という方法である。

②のフィルタにより，抽出対象音節は強調的に抽出される。③のフィルタにより排除対象音節は強調的に排除される。そして⑤により，抽出対象音節はさらに強められる。　他単音節平均フィルタによる抽出音声には，ある程度排除対象音節は残ってしまうが，抽出音声のレベルの小さい高周波成分も通過するため，明瞭度も比較的に良好である。

今後は一つの単音節平均フィルタによって，音程の異なる同一単音節を抽出する方法などについても検討し，適用範囲を広げるための手法を検討したい。

参考文献

[1]青木真理子，古家賢一："騒音下音声強調における空間情報の利用について"，信学技報，EA 2002-11，pp.23〜30，2002年

[2]福田拓章，大西崇浩，東山三樹夫，寺田隆彦，平田能睦："Multi-windowed STFTを用いた基本周波数推定と多声部分離"，信学技報，EA95-99，pp.33〜38，1996年

[3]古井貞熙："音声情報処理"，森北出版，pp.1〜21，1998年

[4]主にプログラミングに関して，小国　力："MATLABと利用の実際"，サイエンス社，1995年

[5]主にプログラミングに関して，小国　力："MathematicaとMATLABの基礎"，サイエンス社，pp.50〜96，1998年

[6]主にグラフィックスに関して，小国　力："MATLABグラフィックス集"，朝倉書店，1997年