第32回信号処理シンポジウム
2017年11月8日〜10日（盛岡）

位相シフトキーイングとしてのバイナリマスク

Binary Masking as Phase-shift Keying

太刀岡勇気
デンソーアイティーラボラトリ

Yuuki TACHIOKA
Denso IT Laboratory

アブストラクト　バイナリマスク法は，時間周波数平面上のスペクトルに対して，各時間周波数ビンでの 2 マイクの位相差から求められる到来方向に基づき，目的音に 1，ノイズに 0 を与えるようなバイナリマスクを構成し，ノイズを抑圧する方法である．これは目的音の存在確率に基づく理想バイナリマスクを，位相差にエンコードし，それを到来方向にデコードしてマスクを推定する問題であると考えると，位相シフトキーイングの問題として，通信理論的解釈を行うことができる．それにより，受信信号からマスクを推定した場合のマスクのエントロピーを信号対雑音比，到来方向，許容角度の関数として数値的に求められることが分かった．音声とガウスノイズを用いた実験により妥当性を確認した．

1　はじめに

騒音がある環境で目的音のみを抽出する技術は様々な応用がある．様々な手法が提案されているが，マイクを複数用いて観測音の到来方向を推定し，目的音の位置情報に基づいて騒音を抑圧する手法が有効である [1]．その中でも，バイナリマスク法は，時間周波数平面上のスペクトルに対して，各時間周波数ビンでの 2 マイクの位相差から求められる到来方向に基づき，目的音に 1，それ以外の方向から到来するノイズに 0 を与えるようなバイナリマスクを構成し，ノイズを抑圧する方法である．これは簡単だが効果的であるため，よく用いられる [2]–[4]．

バイナリマスク法では到来方向や許容角度の条件によって性能が大きく変化するが，それらの影響を理論的に解析した事例は見られないようである．一方で，通信の分野では，送りたいメッセージを位相に変換して通信を行う位相シフトキーイングがよく用いられ，その理論的な解析が行われている [5]．バイナリマスク法を，時間周波数平面上での目的音の有無を表す理想バイナリマスクを，位相差にエンコードして送信し，それを受信後に到来方向にデコードしてマスクを推定する問題であると考えると，位相シフトキーイングの問題として，通信理論的解釈を行うことができる．騒音が存在する場合の観測位相差を，

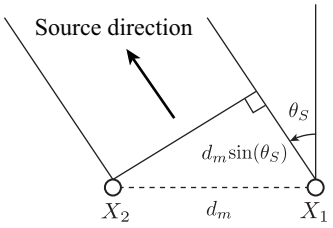

図 1: A microphone setting.

目的音が存在する場合/しない場合に分けてモデリングすることで，受信信号からマスクを推定した場合のマスクのエントロピーを，信号対雑音比，到来方向，許容角度の関数として数値的に求められることが分かった．音声とガウスノイズを用いた実験により妥当性を確認する．

2　バイナリマスク法の「雑音のある離散的通信路モデル」でのモデル化

2.1　バイナリマスク法による騒音抑圧処理

時間周波数平面上でそれぞれのビンが 1bit のマスク情報 M を持っており，対象音が存在するビンは 1，それ以外のビンは 0 で符号化されているものが，理想バイナリマスクであり観測音から推定したい対象であると考える．図 1 のように，ここではマイク 2 本の場合を考え[1]，それらの間隔を d_m[m]，対象周波数を f[Hz]，音速を c[m/s] とする．バイナリマスク法が有効な周波数帯域は，エイリアシングが起こらない $f < \frac{1}{2}\frac{c}{d_m}$ を満たす帯域である．目的音の方向[2]を θ_S[rad]，許容角度を $\theta(>0)$ として，目的音はパスバンド $\theta_p(\theta_S - \theta \le \theta_p \le \theta_S + \theta)$ 内に存在するとして通信が行われ，ノイズが加わって通信される．目的音が存在すれば，マイク間で観測される位相差[3] ϑ_R は，

[1] 3 本以上ある場合は，2 本ずつのペアに対してマスクを構成し，それらを統合することで最終的なマスクを得ることができる．
[2] 正面方向を 0[rad] とする．ただしマイク 2 本では表裏の区別がつかないため，$-\frac{1}{2}\pi \le \theta_S \le \frac{1}{2}\pi$ である．
[3] 到来方向 θ と区別するために，位相差は ϑ で表すことにする．

目的音に由来する位相差

$$\vartheta_S = 2\pi f \frac{d\sin(\theta_S)}{c} = 2\pi \frac{d\sin(\theta_S)}{\lambda} \tag{1}$$

と一致することから，目的音方向 θ_S とマイク間の位相差 ϑ_S が関連付けられる．$\lambda(=c/f)$ は波長である．

ここで，短時間フーリエ変換によるマイク 1 のスペクトルを X_1，マイク 2 のスペクトルを X_2 とすると，それらの位相差は $\vartheta_R = \mathrm{ang}(X_2/X_1)$ で求められる．ここで $\mathrm{ang}(z)$ は複素数 z の偏角 $[0, 2\pi)$ を表す．許容角度 θ を考慮してマスク M は

$$\begin{cases} M = 1: & 2\pi \frac{d\sin(\theta_S-\theta)}{\lambda} \le \vartheta_R \le 2\pi \frac{d\sin(\theta_S+\theta)}{\lambda} \\ M = 0: & \text{otherwise} \end{cases} \tag{2}$$

のように決定される．マスクとスペクトルの積 MX_1 を取り，これを逆短時間フーリエ変換することで騒音抑圧された信号が得られる．

2.2 バイナリマスク法の位相シフトキーイングとしてのモデル化

2.1 節で述べた時間周波数平面上でのバイナリマスク法による騒音抑圧処理を，目的音の有無をマイク間の位相差に変換して通信する問題と考え，「雑音のある離散的通信路モデル」でモデル化する．これは，図 2 のように，信号を位相に変えて通信する位相シフトキーイング (Phase-shift Keying; PSK) と同じであるので，図 3 のように，2 値である時間周波数マスク M を，連続量であるマイクの位相差 ϑ_R に情報を変え，通信する系と考えることができる．

2.3 位相差 ϑ_R の分布の算出法

一般には，観測された位相差 ϑ_R はノイズの影響で，式 (1) とは一致しない．このノイズの影響を定量的に明らかにする．目的音の振幅を S，ノイズの振幅を N とすると，ノイズの位相差を ϑ_N として，

$$\begin{aligned} \vartheta_R &= \mathrm{ang}\left[S\exp(\jmath\vartheta_S) + N\exp(\jmath\vartheta_N)\right] \\ &= \mathrm{ang}\left[\exp(\jmath\vartheta_S) + \frac{N}{S}\exp(\jmath\vartheta_N)\right] \end{aligned} \tag{3}$$

で表される．$\jmath$ は虚数単位である．信号対雑音比 SNR は

$$SNR = 20\log_{10}\left(\frac{S}{N}\right) \tag{4}$$

であることから，これを上式に代入すると，

$$\vartheta_R = \mathrm{ang}\left(\exp(\jmath\vartheta_S) + 10^{-\frac{SNR}{20}}\exp(\jmath\vartheta_N)\right) \tag{5}$$

となる．ここでノイズとして拡散性のものを仮定すれば ϑ_N は $[0, 2\pi)$ で一様分布であるから，ϑ_R の分布は SNR の関数となり，$SNR > 0$ と $SNR \le 0$ で図 4 のように

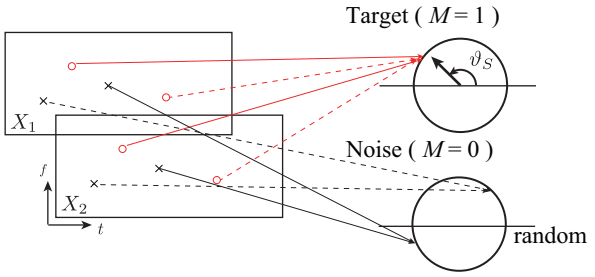

図 2: Conversion of binary masks into phase difference.

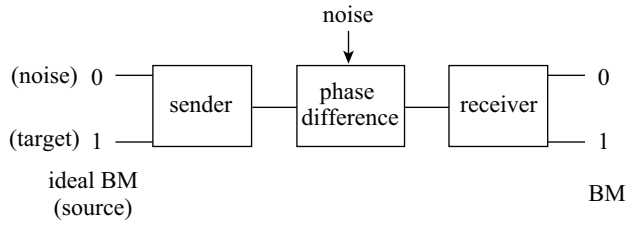

図 3: Communication system sending ideal binary masking to the receiver via phase difference.

ϑ_R の取りうる範囲が異なる．$SNR > 0$ では ϑ_R はある限られた範囲内の値をとるのに対して，$SNR \le 0$ の時には $[0, 2\pi)$ のすべての値をとりうる．$SNR \to \infty$ で ϑ_R は ϑ_S のただ 1 点となり，$SNR \to -\infty$ で ϑ_R は $[0, 2\pi)$ に一様に分布する．よって SNR が与えられれば，ϑ_N が $[0, 2\pi)$ の範囲を動いたときの ϑ_R の変化を考えることで，ϑ_R の確率密度 p_{ϑ_R} を求めることができる．

2.4 信号のエントロピー

ここで送信信号が 0,1 である確率をそれぞれ p_0^s，p_1^s，受信信号が 0,1 である確率をそれぞれ p_0^r，p_1^r とする．また通信理論の慣例に則り，送信信号を x，受信信号を y で表す．まず，送信信号 x のエントロピー

$$H(x) = -p_1^s\log_2(p_1^s) - p_0^s\log_2(p_0^s) \tag{6}$$

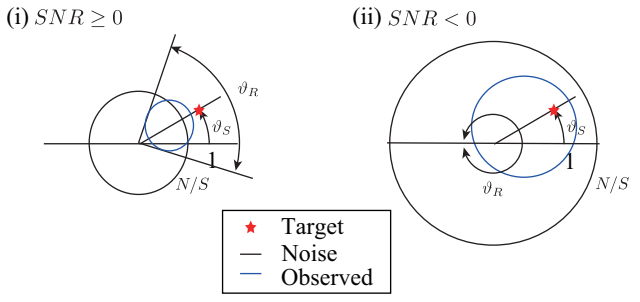

図 4: Range of observed phase difference ϑ_R. If SNR is positive, ϑ_R is limited to the certain range. Otherwise, ϑ_R is distributed over $[0, 2\pi]$.

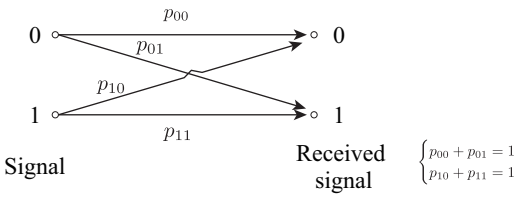

図 5: Discrete channel representing communication system in Fig. 3.

は送信信号の性質のみに依存するため，事前に調べておくことができる．ここから受信信号 y のエントロピー

$$H(y) = -p_1^r \log_2(p_1^r) - p_0^r \log_2(p_0^r) \tag{7}$$

を予測するためには，図5のように，信号 x が送信された際に信号 y が受信される条件付き確率 p_{xy}(=p_{00},p_{01},p_{10},p_{10}) を求める必要がある．これらの確率が求まれば，受信信号の確率

$$\begin{bmatrix} p_0^r \\ p_1^r \end{bmatrix} = \begin{bmatrix} p_{00} & p_{10} \\ p_{01} & p_{11} \end{bmatrix} \begin{bmatrix} p_0^s \\ p_1^s \end{bmatrix} \tag{8}$$

から受信信号のエントロピー (式 (7)) を求めることができる．同様に，条件付きエントロピーは

$$\begin{aligned} H_x(y) = &-p_0^s(p_{01}\log_2 p_{01} + p_{00}\log_2 p_{00}) \\ &- p_1^s(p_{10}\log_2 p_{10} + p_{11}\log_2 p_{11}) \end{aligned} \tag{9}$$

である．

2.5 条件付き確率 p_{xy} の算出法

受信信号のエントロピーを求めるためには，図 3 のシステムを通った時の条件付き確率 p_{xy} を予測できればよい．まず $M=0$ が $M=1$ として受信されてしまう状況を考える．このとき，送信信号 x が騒音であるため ϑ_R は $[0, 2\pi)$ に一様に存在し，これがバイナリマスクのパスバンド θ_p に入るときだから，確率 p_{01} は

$$p_{01} = \frac{\sin(\theta_S+\theta) - \sin(\theta_S-\theta)}{2} \tag{10}$$

$$p_{00} = 1 - p_{01} \tag{11}$$

のようになる．これは SNR によらず，許容角度 θ を大きくすると p_{01} も大きくなる．

次に，$M=1$ が $M=1$ のまま受信できる状況を考える．これは，目的音である送信信号に騒音が加わり，式 (5) が観測され，これが式 (2) の ϑ_R の範囲に入った場合にのみ起こりうる．2.3 節の方法に従って，確率密度分布 p_{ϑ_R} から，式 (2) の $M=1$ の条件を満たす確率を算出すれば，それが p_{11} となる．

$$p_{11} = \int_{\vartheta^-}^{\vartheta^+} p_{\vartheta_R}(\vartheta) d\vartheta \tag{12}$$

$$p_{10} = 1 - p_{11} \tag{13}$$

ここで

$$\begin{cases} \vartheta^+ = 2\pi \frac{d\sin(\theta_S+\theta)}{\lambda} \\ \vartheta^- = 2\pi \frac{d\sin(\theta_S-\theta)}{\lambda} \end{cases} \tag{14}$$

である．p_{11} を大きくするためには，許容角度 θ を大きくすることが有効だが[4]，副作用として p_{01} の確率も大きくなってしまうというトレードオフがあることが解析的に示せる．

3 実験による検討

2 節での解析の妥当性を検証するため，実験的な検討を行った．音声とホワイトノイズを混合し，実験的に条件付き確率 p_{xy} と受信信号のエントロピーを求め，理論値と比較した．音声は WSJ の読み上げコーパス[5]から 1 話者 10 文の読み上げを用意した．クリーン音声の，エネルギーが背景騒音よりも 5dB 以上大きい時間周波数ビンを，理想バイナリマスクのビンとした．原信号と，目的音の到来方向 θ_S に当たる到来時間差を与えた信号の 2 信号からバイナリマスクを構成した．その際，到来方向 θ_S は -60 度，0 度[6]の 2 パターン，SNR は 0,6,12,18[dB] の 4 パターンで実験した．許容角度 θ は 10,20 度の 2 パターンとした．ガウシアンノイズより騒音を作成し，騒音の到来方向がランダムになるよう，2 信号が別々の開始時間となるようにして混合した．

3.1 到来方向 θ_S の影響

目的音の到来方向が，$\theta_S = -60°$ の時と $\theta_S = 0°$ の時の比較をした．許容角度 $\theta = 10°$ で同じとした．$\theta_S = -60°$ の場合の条件付き確率を，図 6 に示す．また受信信号のエントロピーを図 7 に示す．どちらの場合でも，SNR によらず解析値と実験により求めた値はよく一致している．

$\theta_S = 0°$ の場合の条件付き確率を，図 8 に，受信信号のエントロピーを図 9 に示す．$\Delta(\theta_S, \theta) = \sin(\theta_S+\theta) - \sin(\theta_S-\theta)$ は，θ が一定の場合，$\theta_S = 0°$ の時に最大となる．よって，p_{01}，p_{11} のいずれもが $\theta_S = -60°$ の場合よりも大きいため，図 6 と比べて，図 8 では，p_{00} の値が小さくなり，p_{11} の値が大きくなっている．これから到来方向によっても最適となる θ が異なることがわかる．この場

[4] ただし $SNR > 0$ の場合には，ある一定の範囲以上に拡大しても p_{11} は大きくならないことが，図 4 からわかる．

[5] https://catalog.ldc.upenn.edu/LDC93S6B (2017 年 9 月 11 日確認)

[6] 以後はわかりやすさを重視して，角度は [rad] ではなく [度] で表すこととする．

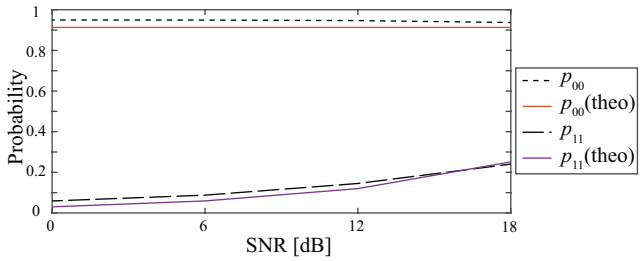

図 6: Conditional probability p_{00} and p_{11} ($\theta_S = -60°$, $\theta = 10°$).

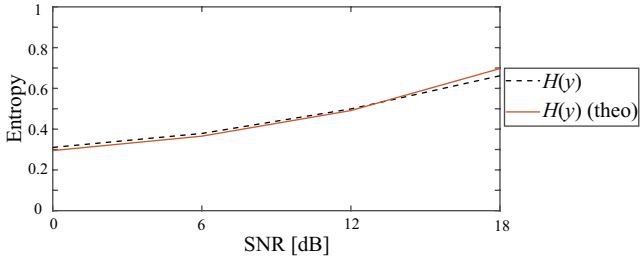

図 7: Entropy of the received signal $H(y)$ ($\theta_S = -60°$, $\theta = 10°$).

合もおおむね傾向は一致しているが，$SNR = 18$[dB] の場合のみ，p_{11} の解析値が実験の値よりも大きくなっている．今回 SNR は発話単位のエネルギーで決定しているが，実際には音声の存在する時間周波数ビンはスパースなため，少数の時間周波数ビンが平均の SNR よりも著しく大きく，それ以外のビンは平均の SNR よりも小さかったため，SNR がよい条件で解析値が p_{11} を大きく見積もったと考えられる．

3.2 許容角度 θ の影響

許容角度の影響を考察するため，θ を 20°，$\theta_S = -60°$ の場合を，図 10，図 11 に示す．θ_S が一定の場合，$\Delta(\theta_S, \theta)$ は，θ が大きくなるほど大きくなる．$\Delta(-60, 10) = 0.174$，$\Delta(0, 10) = 0.347$，$\Delta(-60, 20) = 0.342$ であることから，到来方向 θ_S は図 6，7 と同じであるが，$\Delta(\theta_S, \theta)$ の値の近い図 8，9 と傾向が似ていることがわかる．

4 まとめと今後の課題

騒音を抑圧する手法の 1 つであるバイナリマスク法を，目的音の存在の有無を表す理想バイナリマスクを，位相差により通信しマスクを推定する問題と考えることで，位相シフトキーイングの問題として定式化した．通信理論的解釈により，受信信号のエントロピーを解析的に求められることが分かり，音声とガウスノイズを用いた実験により妥当性を確認した．

今後の課題は，方向性の騒音がある場合にもこの議論

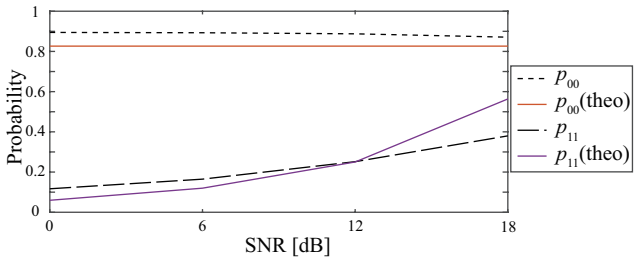

図 8: Conditional probability p_{00} and p_{11} ($\theta_S = 0°$, $\theta = 10°$).

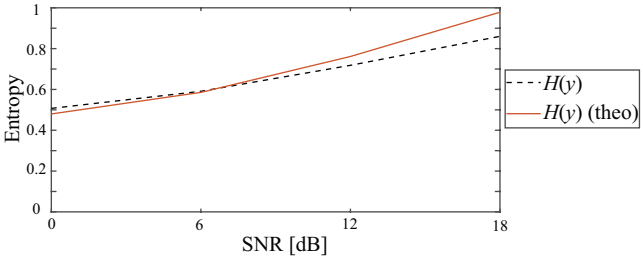

図 9: Entropy of the received signal $H(y)$ ($\theta_S = 0°$, $\theta = 10°$).

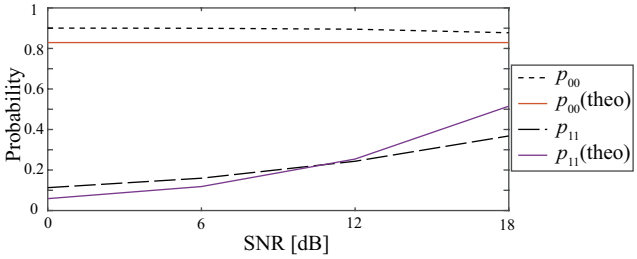

図 10: Conditional probability p_{00} and p_{11} ($\theta_S = -60°$, $\theta = 20°$).

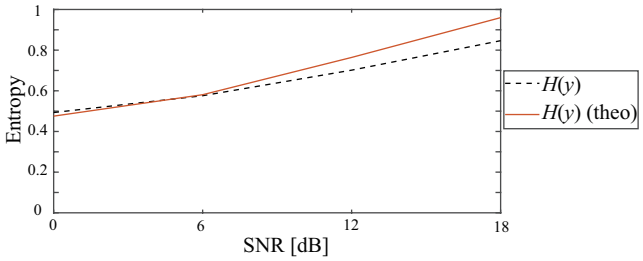

図 11: Entropy of the received signal $H(y)$ ($\theta_S = -60°$, $\theta = 20°$).

を拡張することである．さらにこの分析をもとにした最適な許容角度 θ を与える方法の検討 (例えば 2 節に示した検討により，θ_S により θ を変えるべきであることがわかった) や，残響がある場合の検討などが挙げられる．

参考文献

[1] 浅野太，音のアレイ信号処理 —音源の定位・追跡と分離—，コロナ社 (2011).

[2] H. Sawada, S. Araki, R. Mukai, and S. Makino, “Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation,” IEEE Transactions on Audio, Speech and Language Processing, **15**, 1592–1604 (2007).

[3] J. Cermak, S. Araki, H. Sawada, and S. Makino, “Blind source separation based on a beamformer array and time frequency binary masking,” Proceedings of ICASSP, **1**, pp.145–148 (2007).

[4] S. Liang and W. Liu, “Binary mask estimation for voiced speech segregation using Bayesian method,” Proceedings of Asian Conference on Pattern Recognition, pp.345–349 (2011).

[5] 武部幹，田中公男，橋本秀雄，情報伝送工学，オーム社 (1997).