4T-5

音声 Captcha の自動解析に向けた Single Channel 混合音声の数字認識

佐野 正太郎 † 大塚 琢馬 † 奥乃 博 †

† 京都大学大学院 情報学研究科 知能情報学専攻

1. はじめに

CAPTCHA は人間と機械を判別するためのプログラムであり, 人間には容易で機械には解読不可能な問題を出題する. スパムアカウント登録をはじめとするサービス乱用を防止するため, 多くの Web サービスが CAPTCHA を導入している. 一般的な CAPTCHA は画像上に複数の文字を表示し, それらを読みとれるか否かにより人間と機械を判別する. 一方で, いくつかの Web サービスはアクセシビリティを考慮し, 音声聞き取り問題による CAPTCHA を提供している.

先行研究 [6] [1] [5] はいくつかの音声 CAPTCHA を提供するサービスを対象とし, それらを自動解読することのできるシステムを構成することで音声 CAPTCHA の防御手法を評価している. 一般的な音声 CAPTCHA では, ユーザは複数のオーバーラップしないターゲット音声 (数字, アルファベット, 単語など) を聞き取るが, ターゲット音声には別話者音声・同一話者音声などの妨害音が混合されている. Stilt Warker [5] により, ターゲット音声のセグメント化が容易な場合には, このような CAPTCHA が容易に自動解読されることが示されている.

近来, Google 社の reCAPTCHA は音声 CAPTCHA のための新たな防御手法を導入しており, ターゲット音声をランダムにオーバーラップさせることで機械によるセグメント化を防止している. 本稿ではランダムにオーバーラップしたターゲット音声をセグメント化できるシステムを提案する.

2. 音声 reCAPTCHA のスキーマ

図 1 に音声 reCAPTCHA の一例を波形で示す. 本稿では音声 reCAPTCHA の 1 問題を “チャレンジ” と呼ぶ. チャレンジは 3 つの “クラスタ” からなり, 1 つのクラスタは 3 または 4 個の “ディジット” を含んでいる. クラスタ内のディジットはランダムな間隔で互いにオーバーラップしている. チャレンジ内の全てのディジットが正しく入力されたとき, reCAPTCHA は入力者を人間であるとみなす. 要約すると, 機械によるセグメント化に対し 2 つの対策を講じている: (1) ターゲット音声どうしをランダムな間隔でオーバーラップさせる, (2) ターゲット音声の数をランダムに与える.

3. reCAPTCHA ソルバー

図 2 に我々のシステムの概要を示す. 我々のシステムは音声 reCAPTCHA のチャレンジ 1 つを入力とし, その解答を出力する. システムは cluster segmentation, digit database (DB), MFCC extraction, MFCC classification, digit segmentation, digit classification の 6 つのコンポーネントからなる. システムは以下の手順でチャレンジを解読する: (1) チャレンジ音声を 3 つのクラスタ音声に分解する (cluster segmentation), (2) クラスタ音声をディジット音声にセグメント化する (digit segmentation), (3) ディジット音声をラベル付けする (digit classification).

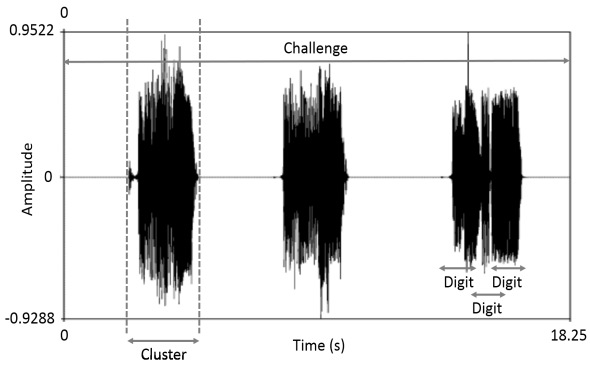

図 1: Audio reCAPTCHA

Digit segmentation と digit classification では入力音声の Mel-frequency cepstral coefficients (MFCC) ベクトルの抽出と, 抽出ベクトルのクラス分類が行われる (MFCC extraction, MFCC classification). Digit DB は MFCC classification の学習データとなるラベル付き数字発話音声を格納している.

3.1 Cluster Segmentation

このコンポーネントは 1 つのチャレンジ音声を 3 つのクラスタへと分解する. クラスタ間は無音であるため, 以下に示すような音量に基づく有音区間検出アルゴリズムを適用する: (1) 音声全体を固定区間の窓に分割する, (2) それぞれの窓に対して信号の分散を計算する, (3) 区間内の全ての窓の分散が一定値を超える最長区間を 3 つ抽出する.

3.2 Digit DB

Digit DB には MFCC classifier の学習に使われるラベル付き数字発話音声が格納されている. それぞれのラベル $(0,\ldots,9)$ に対して 150 個, 合計で 1500 個の音声が格納されている. これらの音声は実際の音声 reCAPTCHA から手動でセグメントおよびラベル付けされたものであり, 2012 年 11 月にダウンロードされた.

3.3 MFCC Extraction

入力音声の短時間窓ごとに MFCC ベクトルを抽出する. 我々は窓幅を 25ms とし, 13 次元 MFCC ベクトルを利用した.

3.4 MFCC Classification

このコンポーネントは MFCC ベクトル f を入力とし, f の各ラベルに対する確率 $p_0(f),\ldots,p_9(f)$ を出力する. このコンポーネントの実体は, 多クラスの確率推定に拡張 [4] された Support Vector Machine (SVM) である. 以下の手順により, SVM の学習を行った: (1) digit DB の音声それぞれを MFCC extraction に入力する, (2) ステップ

How to Break Overlapped Audio CAPTCHA: *Shotaro Sano (Kyoto Univ.), Takuma Otsuka (Kyoto Univ), and Hiroshi G. Okuno (Kyoto Univ.)*

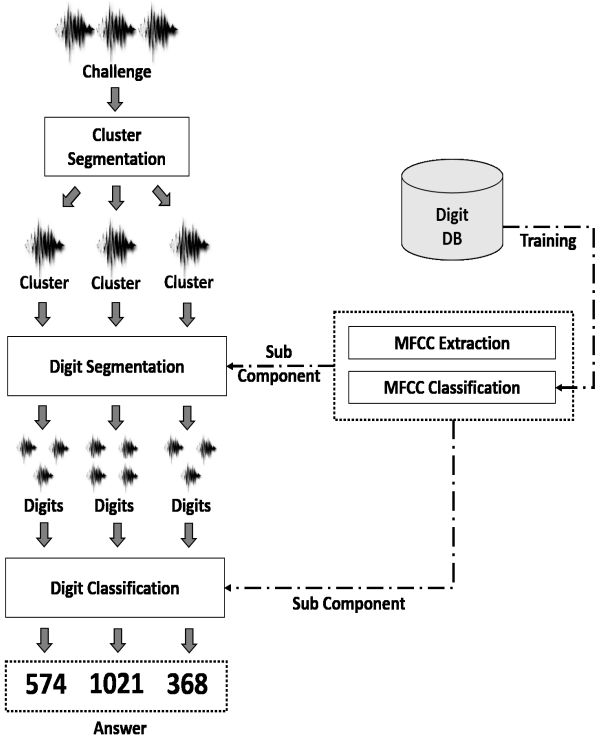

図 2: Solver Overview

1 で抽出される MFCC ベクトルそれぞれに対して抽出元音声のラベルを付与する, (3) ステップ 2 で得られるラベル付き MFCC ベクトルを SVM の学習データとする.

3.5 Digit Segmentation

Digit segmentation への入力は 1 つのクラスタ音声である. このコンポーネントは入力クラスタ音声を漏れなく重複なく 3 または 4 つのディジット音声へとセグメント化する.

3.5.1 問題設定

図 3 に示すように, 時間軸を MFCC 窓に対応させて区切る. 時刻 0 はクラスタの先頭に, 時刻 w は末尾に相当する. 時刻 A_s から B_s にまたがるセグメント s を $s=[A_s,B_s]$ のように表す.

最適なセグメンテーション $s_0,\ldots,s_N$ は次式に示す目的関数 G_N^w を最適化する:

$$G_N^w(s_1,\ldots,s_N)=\max_{N,s_1,\ldots,s_N}\sum_{i=1}^{N}\frac{g(s_i)}{N} \tag{1}$$

$g(s)$ はセグメント s のスコアであり, s から抽出される MFCC ベクトルに対する SVM スコアの平均を用いて次式のように定義される:

$$g(s)=\max_{l\in\{0,\ldots,9\}}\frac{1}{B_s-A_s}\sum_{t=A_s}^{B_s-1}p_l(f(t,t+1)), \tag{2}$$

$f(t,t+1)$ は時刻 t から $t+1$ で抽出される MFCC ベクトルである. 更に, 次の拘束条件が加わる: (a) $N=3\cup N=4$, (b) $A_{S_1}=0$, (c) $B_{S_N}=w$, (d) $B_{S_i}=A_{S_{i+1}}(i=1,\ldots,N-1)$.

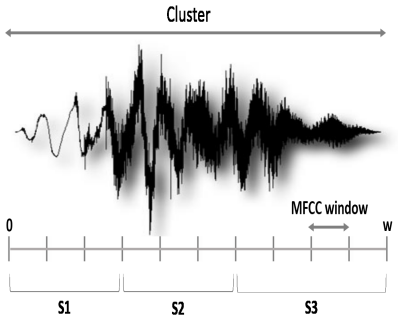

図 3: Digit Segmentation

3.5.2 解法

目的関数 G_N^w の最適解は部分音声のセグメント化とスコアリングを段階的に繰り返す動的計画法によって得られる. 解法の詳細は画像 CAPTCHA に対するセグメント手法を提案している先行研究 [3] に記述されている.

3.6 Digit Classification

このコンポーネントはディジット音声を入力とし, そのラベルを出力する. ラベルは SVM スコアの平均を用いて次式により決定される:

$$\arg\max_{l\in(0,\ldots,9)}\frac{1}{B_s-A_s}\sum_{t=A_s}^{B_s-1}p_l(f(t,t+1)), \tag{3}$$

A_s と B_s はディジットの先頭と末尾に対応する.

4. 実験

実際の音声 reCAPTCHA のデータを用いてシステムの性能を評価した. 2012 年 11 月時点の音声 reCAPTCHA から 100 個のチャレンジをダウンロードし, システムに入力した. 結果として, システムは 100 個中 12 個のチャレンジに正解した.

5. まとめ

我々のシステムは 12%の精度で音声 reCAPTCHA を自動解読した. “サービスの規模と自動解読のコストにもよるが, $\frac{1}{10000}$ 以上の精度を持つ自動解読システムは危険である”[2] ことを考えると, この結果は音声 reCAPTCHA にみられる防御手法の脆弱性を示したと言える.

参考文献

[1] E. Bursztein, R. Beauxis, H. Paskov, D. Perito, C. Fabry, and J. Mitchell. The failure of noise-based non-continuous audio captchas. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pages 19–31. IEEE, 2011.

[2] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski. Building segmentation based human-friendly human interaction proofs (hips). *Human Interactive Proofs*, pages 173–185, 2005.

[3] J.B. Fiot and R. Paucher. The captchacker project. *Ecole Centrale Paris*, 2009.

[4] T.K. Huang, R.C. Weng, and C.J. Lin. Generalized bradley-terry models and multi-class probability estimates. *The Journal of Machine Learning Research*, 7:85–115, 2006.

[5] Stiltwalker. recaptcha v1. `http://www.dc949.org/projects/stiltwalker`.

[6] J. Tam, J. Simsa, S. Hyde, and L. Von Ahn. Breaking audio captchas. *Advances in Neural Information Processing Systems*, 1(4), 2008.