2003－SLP－47　(14)
2003／7／19

タッピングを利用した音声認識の検討

番 弘光[*]　伊藤 克亘[†]　武田 一哉[†]　板倉 文忠[‡]

[*‡] 名古屋大学大学院工学研究科
[†] 名古屋大学大学院情報科学研究科

〒 464-8603 名古屋市千種区不老町 1
[*]hban@itakura.nuee.nagoya-u.ac.jp　[†]{itou,takeda}@is.nagoya-u.ac.jp
[‡]itakura@nuee.nagoya-u.ac.jp

あらまし　本研究では，連続数字音声認識において音声情報に加え，タッピングにより与えられるタイミング情報を利用した音声認識を提案する. 具体的には，タッピングした回数により，記述文法を切り替えることによって連続数字の桁数を事後的に決定する方法である. また，タッピングのタイミングから単語境界を推定し，音声認識に利用する手法についても検討し，汎用連続音声認識パーザ Julian に提案手法の実装を行った. これらの手法を雑音を重畳したデータベースに適用し，その性能評価を行った結果，最大で 48.54% の改善率を得ることができた.

A study on speech recognition using tapping

Hiromitsu BAN[*], Katsunobu ITOU[‡], Kazuya TAKEDA[‡]
and Fumitada ITAKURA[†]

[*‡] Graduate School of Engineering, Nagoya University
[†] Graduate School of Information Science, Nagoya University

Furo-cho 1, Chikusa-ku, Nagoya 464-8603, JAPAN
[*]hban@itakura.nuee.nagoya-u.ac.jp　[†]{itou,takeda}@is.nagoya-u.ac.jp
[‡]itakura@nuee.nagoya-u.ac.jp

Abstract　In this paper, speech recognition method using the time information given by tapping for the connected digit recognition is proposed. On this method, the number of digits is decided by tapping times for changing grammar, and word boundary information is estimated, then we combine them. Our method is ported to a general-purpose speech recognition parser Julian. In evaluation on noisy speech database, maximum 48.54% relative implovement rate is obtained using our method.

1　はじめに

確率・統計的手法により，現在の音声認識システムの認識性能は格段に向上している．しかしながら，実環境においては，様々な背景雑音や室内残響の影響などにより，認識性能が劣化するという問題があり，このことが音声認識の実用化の妨げとなっている．

雑音に頑健な音声認識システム確立のためのアプローチとして，雑音除去が挙げられ，従来スペクトルサブトラクション法 (以下 SS 法)[1] がよく用いられている．SS 法は雑音が重畳した音声のパワースペクトルから、推定された雑音のパワースペクトルを減算することにより，音声のパワースペクトルを推定する手法である．SS 法は，演算量が少なく効果的な方法として広く採用されている．

しかし，条件の悪い雑音環境下では音響的な情報だけでは限界がある．そこで本研究では，音声と同時に使えるモダリティとしてタッピングを利用することを提案する．タッピングとはデバイス（マウスのボタンやキーボードのキー）を押したり叩いたりして回数やタイミング情報を入力するインタフェースである．本研究では，タッピングによって，タイミング情報及び回数を音声認識に利用する方法を提案する．タッピングのタイミングと音声の例を図 1 に示す．

本研究では，提案手法であるタッピングを利用した音声認識の実装及び評価のため，タッピングとともに雑音下音声認識評価用データベースである AURORA2[2] を日本語に翻訳した音声データを収集し，提案手法の評価を行った．

2　雑音下音声認識評価用データベースの作成

本研究で作成及び使用したデータベースについて概説する．

2.1　話者と読み上げ文

収集したデータベースの発話内容は，AURORA2 で用いられている各連続数字をそのまま日本語に翻訳したものとした (AURORA2J[3])．データベースの連続数字は 1~7 桁 (6 桁は除く) で，桁数の平均は 3.29 である．音声及びタッピングの収録を行った話者数は 214 名 (男女各 107 名) である．

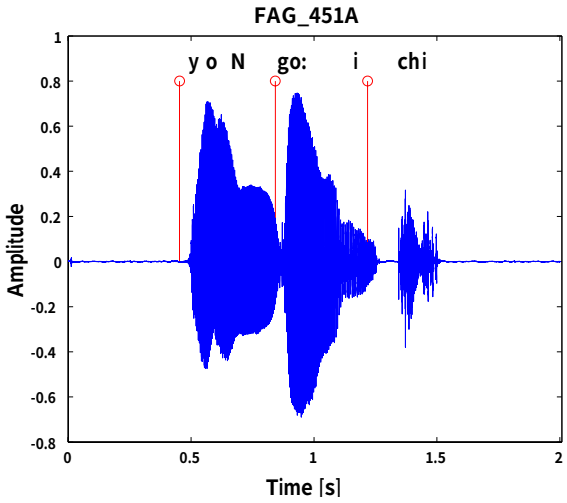

図 1: タッピングと音声の例（離散プロットがタッピングのタイミングを示している）

2.2　音声及びタッピングの収録

収録機材として，計算機に Sony VAIO PCV-MXS2，A/D 変換器に EDIROL UA-5，マイクに Sennheiser HMD410 を用いて，防音室にて行った．被験者は机上に置かれた数字の書かれた原稿を読み，その音声の録音を行った．タッピングのデバイスとしてキーボードを採用し，キーボードを机上に置き，被験者がキー ([Enter] キー) を押下することによってタイミング情報を収録した．音声の標本化周波数は 16kHz である．収録の様子を図 2 に示す．

2.3　雑音下音声認識評価用データベース

収録した音声を計算機上で種々の雑音を重畳することにより雑音下音声認識評価用データベースを生成した．重畳した雑音と SNR，伝達特性は AU-

図 2: 音声録音の様子

RORA2 と同一である. テストセットには表 1 に示す，3 種類のセットを用意した.

表 1 において，SetA，SetB ではそれぞれ 4 種類，SetC では SetA，SetB から 1 種類ずつ選択した雑音が用いられ，SNR は −5 ～ 20dB(5dB 刻み) 及びクリーン環境を用意した. また，各雑音，SNR ごとに 1001 文章の音声データ (男女混在) が含まれており，各音声データの標本化周波数は 8kHz(16bit) にダウンサンプリングしている.

HMM の学習セットとしては，クリーン音声のみの学習データセット (Clean Training Condition) と，雑音重畳音声を含んだ学習データセット (Multi Training Condition) の 2 種類の学習データセットを用意した. Multi Training Condition に含まれる雑音重畳音声データには，テストセット SetA に含まれる 4 種類の雑音が重畳しており，SNR はクリーン及び，0 ～ 20dB のみである.

認識実験の枠組みも AURORA2 と同一で，クリーンな音声により学習された HMM と，雑音重畳音声により学習された HMM それぞれを用いて評価を行う. 評価は単語誤り率及び，ベースラインの単語誤り率に対する改善率により行う. 実質的な評価に用いたのは，各テストセットの SNR0～20dB における単語誤り率，改善率の平均値であり，クリーン及び-5dB 環境は評価の対象とはしない. ここで，改善率は以下の式で表される.

$$\frac{\text{提案手法の認識率} - \text{ベースラインの認識率}}{100 - \text{ベースラインの認識率}} \quad (1)$$

HMM の学習は HTK(Hidden Markov Model Toolkit)[4] により行った. 認識時の語彙数は 13(数字 1～9，maru，zero，無音，ショートポーズ) であり，各語ごとに HMM を学習した. 学習した HMM の構造は表 2 の通りであり，ショートポーズの HMM は無音 HMM の第 3 状態を共有している.

デコーダには汎用連続音声認識パーザ Julian3.3[7] を用いており，認識時には第 1 パスのみを用いた.

表 1: データベースに含まれる雑音環境

セット名	雑音環境	フィルタ特性
SetA	Subway，Babble Car，Exibition	G.712
SetB	Restaurant，Street Airport，Station	G.712
SetC	Subway，Street	MIRS

表 2: 学習した HMM の構造

	状態数	混合分布数
数字 (1～9，maru，zero)	16 状態	20
無音	3 状態	36
ショートポーズ	1 状態	36

表 3: 音響分析条件

標本化周波数	8kHz
フレーム長	25ms
フレーム周期	10ms
時間窓	ハミング窓
特徴量	12 次 MFCC + Log-Power + 12 次 ΔMFCC + ΔLog-Power + 12 次 ΔΔMFCC + ΔΔLog-Power

ベースラインの認識では文法 (言語モデル) は数字が任意の回数繰り返すという記述文法を用い，ビーム幅は 100，挿入ペナルティは 0 とした.

HMM の学習及び認識時の音響分析条件は表 3 の通りである.

本研究で作成した雑音下音声認識評価用データベースのベースラインの認識結果を表 4 に示す. 本研究では，以上のようなデータベースを用いて，提案手法の評価を行っている.

表 4: ベースラインの認識結果

Reference Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	6.62%	8.81%	11.78%	8.53%
Clean	47.95%	47.09%	48.47%	47.71%
Average	27.29%	27.95%	30.13%	28.12%

3 タッピングを利用した音声認識

3.1 タッピングの回数による桁数の決定

タッピングの利用方法として，タッピングした回数によって連続数字の桁数を事後的に決定するという方法が挙げられる. 具体的な処理として，認識時に数字が任意の桁数繰り返すという文法ではなく，タッピングした回数だけ数字が繰り返すという文法に切り替えるという手法である. この処理により，タッピングの回数が正しければ連続数字の桁数が正

しく認識され，認識性能の改善が望める．

この処理では，タッピングの回数が誤っていれば，必ず誤認識がおこるという問題がある．しかし，評価用セット作成のために収集した 4004 文でタッピングの回数を誤ったのは 14 文と少なく，タッピングの回数を誤ることによる認識性能への影響は軽微だと考えられる．

3.2 単語境界とタッピングのずれ

雑音環境下で頑健に音声認識を行う手法として，韻律情報を利用する手法が研究されている [5]. 特に，基本周波数 (F_0) パターンは，句や単語境界の推定に役立つとされ，雑音重畳音声の認識性能向上に有効であることが示唆されている [6].

そこで，タッピングのタイミングと単語境界のずれの分布を統計的に調査し，単語境界を推定することを考える. 学習セット 8440 文中のタッピングの回数が正しい 8431 文 (27689 単語) に対して単語境界の推定を行い，タッピングと単語境界のずれを算出した. 単語境界の推定は，学習された HMM に対して，最も高い確率を与える状態遷移系列により決定する、いわゆる forced alignment 法による。アライメントを行ったデータは学習データに対して Closed である．

全単語の単語境界とタッピング時間差の平均は −27.65ms(単語の発声の方がタッピングよりも 27.65ms 遅い)，標準偏差は 88.28ms であった. 単語境界とタッピングのずれを図 3 に示す. フレーム数が正のときにタッピングの方が単語の発声よりも遅れていることを示す. 図 3 より，フレームのずれはおよそ 20 フレーム (200ms) 程度に収まっていることがわかる. また，92.1%の単語 (25511) が± 15 フレーム (±150ms) に，73.2%の単語 (20280 単語) が± 10 フレーム (±100ms) の範囲に収まった．

3.3 リスコアリングを用いたタッピングのタイミングと音声認識の統合法

図 3 より，タッピングされた時間からある範囲内で単語間遷移が起きる可能性が高いことが確認された. そのため，提案手法としてある範囲内で単語間遷移が起きなかった仮説のスコアに対してはペナルティを与えることによってリスコアリングを行う手法を提案する. この操作により，音声のアライメントが正しくない可能性が高い仮説の尤度を下げることができると考えられ，認識性能の向上が期待できる. 今回の報告では，設定した範囲外で単語間遷移

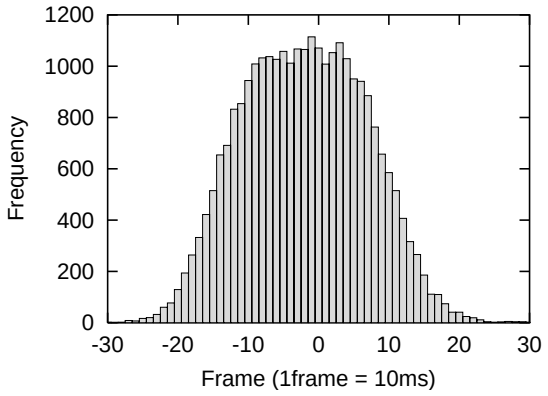

図 3: タッピングと単語境界のずれ

が起きた仮説には一律にペナルティを与えている．

4 スペクトルサブトラクション法

スペクトルサブトラクション法 [1] は，非音声区間の信号より雑音の特徴量を推定しておき，雑音重畳音声の特徴量から雑音の特徴量を差し引くことにより，元の音声信号を推定する手法である．

音声と雑音が統計的に独立で，雑音が定常的であると仮定すると，雑音付加音声のパワースペクトル $Y(f)$ は音声のパワースペクトル $X(f)$ と雑音のパワースペクトル $\hat{N}(f)$ の和となることから，音声パワースペクトル $\hat{X}(f)$ を次式で推定する．

$$\hat{X}(f) = \max\{Y(f) - \alpha\hat{N}(f), N_f(f)\} \qquad (2)$$

ここで，α はサブトラクション係数，$N_f(f)$ はフロアリングのパワースペクトルである．

5 認識実験

5.1 タッピングにより与えられたタイミングを利用した音声認識

タッピングによる単語境界の推定する手法を汎用連続音声認識パーザ Julian[7] のバージョン 3.3 の第 1 パス (時間同期) に実装し評価を行った. 遷移制限を課すフレームの幅を変えたときに、認識性能がどのように変化するかを分析するために，それぞれの認識結果 (clean，−5dB 環境は除く) の置換誤り，挿入誤り，脱落誤りの集計を行った. 集計を行った単語の総数は，両環境それぞれ 164415 である. ペナルティの値は実験的に 20 と決定した. 音響分析条件は表 3 の条件に従っており，実験条件はベースラインの認識と同一である．

提案手法の認識誤りの内訳を図 4 に示す. 図 4 よ

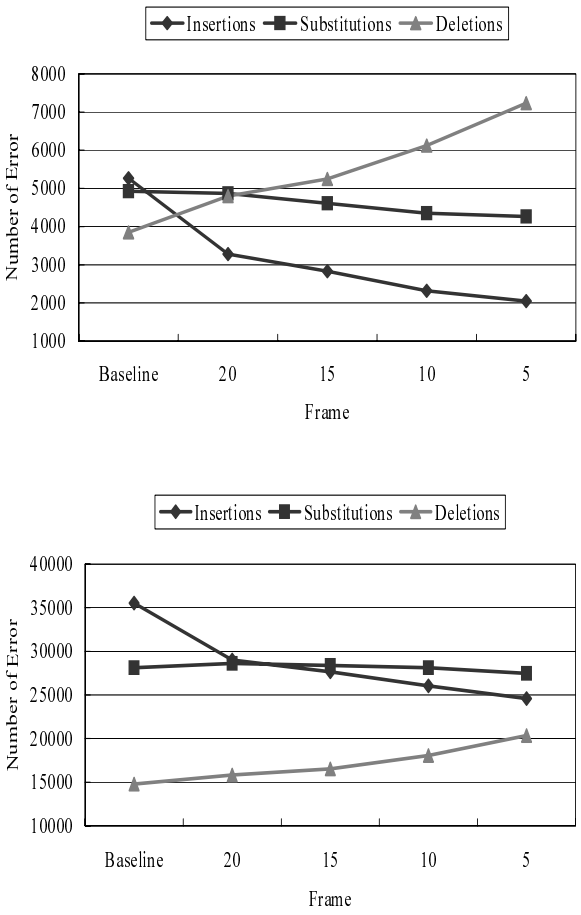

図 4: 認識誤りの内訳 (上:Multi Training Condition, 下:Clean Training Condition)

り，Multi Training Condition，Clean Training Condition，両環境においてペナルティを与える幅を狭くすると置換誤り，挿入誤りは減少し，削除誤りが増加する傾向が見られた. この理由として，ペナルティを与える幅を狭くするほどシステム最適化パラメータである挿入ペナルティの役割に近づいていき，挿入誤りが抑制されるとともに脱落誤りが増加したと考えられる.

5.2 種々の手法の組合わせによる認識実験

提案手法の有効性を確認するために SS 法を雑音下音声認識評価用データベースに適用して評価を行った. また，提案手法及び SS 法を組合わせた場合の認識性能も調査した. それぞれの HMM を用いた認識において，以下の四つの方法で評価を行った.

手法 1 :タッピングによる単語境界の推定

手法 2 :タッピングの回数による文法の切り替え

手法 3 :手法 1 + 手法 2

手法 4 :手法 1 + 手法 2 + SS

手法 1 においてフレームの幅を ±10 フレーム，±15 フレームにした場合については誤りの総数はほとんど変わらなかったが，±15 フレームの場合に最も高い改善率 (両環境の平均) が得られたため，実験では ±15 フレームを選択した. SS 法のパラメータは実験的に $\alpha = 1.2$，$N_f(f) = 0.95$ と決定した.

5.3 実験結果

提案手法及び SS 法による認識結果を表 5-7 に示す. SS 法の平均の改善率は 29.31%，手法 2 の改善率は 31.00% となり，手法 2 の有効性が示された. また，手法 1 の改善率は 17.99% となり，手法 2 による改善率を大きく下回った. この理由として，時間差を算出する際に，単語境界は HMM を用いて音響モデルを学習することによって行っているため，その際の単語境界の誤差を含んでいる可能性がある. このため，手法 1 の改善率が手法 2 の改善率を下回ったと考えられる. また，本手法では，タッピングによるリスコアリングの際に，条件から外れた仮説に対して一律にペナルティを与えている. しかし，図 3 からタッピングのタイミングと単語境界とのずれはある分布に従うと考えられ，リスコアリング時にタッピングのペナルティを図 3 を関数近似して与えることによって，さらに認識性能が改善される可能性があると考えられる.

提案手法及び SS 法の組合わせによる手法 3，4 の認識結果を表 8，9 に示す. 手法 3 の平均の改善率は 34.28%，手法 4 の改善率は 48.54% となり，それぞれの手法を単独で用いるよりも高い改善率が得られ，特に全ての手法を組合わせた手法 4 で最も認識性能が改善されることが確認された.

6 まとめ

本研究では，タッピングを利用した音声認識について提案し，その評価のために雑音下音声認識評価用データベースの作成を行った. 提案手法を連続音声認識パーザ Julian に実装し，作成した雑音下音声認識評価用データベースを用いて評価した結果，最大で 48.54% の改善率が得られた. しかし，タッピングという行為をユーザに課すという制限の代償に値するような大幅な改善は得られず，実用化には十分でないと考えられる.

今後の課題として，タッピングよるペナルティの値の最適化及びペナルティを与えるフレームの幅に対する詳細な検討を行う予定である. また，提案手

法を実装した Julian のビーム幅を変化させ，認識性能及び処理時間の変化について調査を行う必要がある．

参考文献

[1] S.Boll，”Suppression of acoustic noise in speech using spectral subtraction”. IEEE Trans. ASSP, vol.ASSP-27，no.2，pp.113-120，1979.

[2] H.G. Hirsch and D. Pearce，”The Aurora2 experimental framework for the performance evaluation of speech recognition systems under noisy conditions”. ISCA ITRW ASR 2000，Sep. 2000.

[3] 中村哲，武田一哉，黒岩真吾，山田武志，北岡教英，山本一公，西浦敬信，藤本雅清，水町光徳．”SLP 雑音下音声認識評価のための WG:評価データ収集について”．情報処理学会研究報告，SLP45-9，pp.51-55，2003.

[4] S. Young，D. Kershaw，J. Odell，”The HTK book”. Entropic，1999.

[5] 岩野 公司，関 高浩，古井 貞熙，”雑音に頑健な基本周波数抽出法とその音声認識への適用”. 電子情報通信学会技術研究報告，SP2002-13，vol.102，no.35，pp.37-42 (2002-4).

[6] 川中 真護，中井 満，下平 博，”F_0 生成モデルに基づくピッチパターン整合を用いた雑音重畳単語音声の認識”. 春季音講論，3-6-14，pp.109-110.

[7] 李昇伸，河原達也，堂下修司，”文法カテゴリ対制約を用いた A*探索に基づく大語彙連続音声認識パーザ”. 情報処理学会論文誌，Vol. 40，No. 4，pp. 1374-1382， 1999.

表 5: 手法 1 による認識結果

Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	7.26%	7.13%	9.75%	**7.71%**
Clean	44.28%	43.05%	46.00%	**44.13%**
Average	**25.77%**	**25.09%**	**27.88%**	**25.92%**

Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	10.79%	35.97%	23.83%	**23.47%**
Clean	12.20%	14.20%	9.70%	**12.50%**
Average	**11.50%**	**25.08%**	**16.77%**	**17.99%**

表 6: 手法 2 による認識結果

Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	5.51%	5.51%	7.66%	**5.94%**
Clean	40.52%	38.82%	42.20%	**40.18%**
Average	**23.01%**	**22.16%**	**24.93%**	**23.06%**

Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	20.07%	50.21%	35.63%	**35.24%**
Clean	25.86%	29.39%	23.30%	**26.76%**
Average	**22.97%**	**39.80%**	**29.46%**	**31.00%**

表 7: SS 法による認識結果

Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	5.55%	6.00%	5.65%	**5.75%**
Clean	37.65%	33.56%	29.74%	**34.43%**
Average	**21.60%**	**19.78%**	**17.70%**	**20.09%**

Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	20.58%	46.27%	36.39%	**34.02%**
Clean	10.71%	28.19%	45.21%	**24.60%**
Average	**15.65%**	**37.23%**	**40.80%**	**29.31%**

表 8: 手法 3 による認識結果

Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	5.88%	5.38%	7.17%	**5.94%**
Clean	39.39%	37.63%	41.42%	**39.09%**
Average	**22.63%**	**21.50%**	**24.30%**	**22.52%**

Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	19.21%	53.38%	41.46%	**37.33%**
Clean	31.18%	33.82%	26.15%	**31.23%**
Average	**25.20%**	**43.60%**	**33.80%**	**34.28%**

表 9: 手法 4 による認識結果

Word Error Rate				
	Set A	Set B	Set C	Overall
Multi	4.82%	4.19%	4.37%	**4.48%**
Clean	27.93%	24.50%	24.65%	**25.90%**
Average	**16.38%**	**14.34%**	**14.51%**	**15.19%**

Relative Improvement				
	Set A	Set B	Set C	Overall
Multi	27.12%	57.88%	44.26%	**42.85%**
Clean	49.44%	56.43%	59.36%	**54.22%**
Average	**38.28%**	**57.16%**	**51.81%**	**48.54%**