Vol.2016-SLP-112 No.10
2016/7/29

音声認識におけるフレームシフト再考

伊藤彰則[†1]

概要　音声認識の特徴量抽出では，まず入力音声に時間窓をかけて局所的な信号を切り出し，音声信号の局所的な周波数情報を取り出す．この時間窓の位置を少しずつずらしながら分析を行うことで，音声の持つ周波数成分の時間変化を捉えることができる．このとき時間窓をずらす時間間隔がフレームシフトであり，典型的には 5ms～10ms に設定される．本稿では，このフレームシフトを 2 つの点から再考する．一つ目の視点は「フレームシフトは 10ms で十分なのか？」という点である．フレームに基づく処理は，音声信号が短い時間で大きく変化しないことを前提としているが，破裂子音などではこの前提がそもそも成立していない．そこで，10ms ごとのフレームの先頭位置のずれによって，抽出される特徴量が大きく変化することを実験的に示す．また，偶然によるフレーム位置の変動に起因する特徴量変動に対応するため，フレーム位置をずらした学習サンプルを学習に用いる方法を提案する．二つ目の視点は，「フレーム位置のずれが学習によって吸収できるのであれば，フレームシフトはもっと長くてもよいのではないか？」という点である．フレームシフトを実験的に 60ms 程度まで長くして実験を行ったところ，フレームシフト 40ms ではフレームシフト 10ms を越える認識性能が得られ，50ms でも 10ms と同程度の認識性能が得られた．これらの条件では 1 状態 HMM（すなわち GMM）が使われており，認識のための計算量の大幅な削減が期待できる．

Reconsidering Frame Shift in Speech Recognition

AKINORI ITO[†1]

Abstract During the feature extraction process for speech recognition, an window function is first applied to the input waveform to extract temporally-limited spectrum. By shifting the window function with a short time period, we can analyze temporal change of speech spectrum. This time period is called “the frame shift”, which is usually 5 to 10 ms. In this paper, frame shift is re-considered from two aspects. The first one is appropriateness of 10 ms as the frame shift. The frame-based process is based on an assumption that temporal change of speech spectrum is slow enough compared with the frame shift, which does not hold for kinds of consonants such as plosives. Thus I experimentally shows that feature value fluctuates much according to the first position of the frame. Then a training method is proposed that uses temporally shifted samples as independent samples to compensate fluctuation of feature caused by the difference of beginning position of a frame. The second aspect is that the frame shift could be longer if the fluctuation can be compensated. To prove this, an experiment was conducted to change frame shift from 10 to 60 ms, and it was found that the result of 40ms frame shift outperformed the result of 10 ms frame shift, and comparable recognition performance with 10 ms frame shift result was obtained with 50 ms frame shift.

1. はじめに

自動音声認識を行うためには,最初に入力音声を分析し，ほとんどの場合は周波数領域の特徴量を抽出する．このためには，入力音声に対して適当な長さの時間窓をかけ，その時間窓の中の音声サンプルに対して FFT や LPC 分析などの変換を行う．この分析結果から，認識に用いる特徴量を計算する．時間窓の位置をずらしながらこの処理を繰り返すことによって，入力音声を時間－周波数の２次元の情報を持つ特徴量に変換することが可能になる．この時に時間窓をずらす時間幅がフレームシフトであり，通常は 5～20ms 程度（典型的には 10ms）の固定値をとる [1]．この時の分析窓幅は 20～25ms 程度である．

本稿では，音声認識のためのフレームに基づく音声分析において，これまで問題にされてこなかった二つの疑問点を再考する．一つは，「フレーム分析は 10ms 程度の離散的分析でよいのか」という問いである．通常のフレームに基づく音声分析では，入力音声信号の 1 サンプル目を起点として，一定のフレームシフトごとに分析窓を設定して分析を行う．そのため，ある音素があった場合に，その音素のどこに分析窓をかけて分析するのかは偶然によって決まる．そのため，破裂子音のような非定常な音素の場合には，窓関数の位置によって特徴量が大きく変わることが考えられるが，音声認識手法の中で，このような「分析窓の位置の変動」に対応する手法は提案されたことがないように思われる．これに対して，本稿ではフレーム開始位置をずらした音声を独立した音声サンプルとして学習に加える方法を検討し，この方法によって分析窓位置のずれによる変動が保証できることを示す．

二つ目の疑問は，「フレームシフトは 10ms 以下である必要があるのか」という問いである．音声分析の目的が「音素の時間周波数構造を詳細に分析する」ことであれば，音素継続長よりも十分短い時間間隔で分析を行うことには合理性がある．しかし，現在の音声認識においては，音素環境に依存する単位（triphone など）を丸ごとモデル化することが主な目的であり，Δ特徴量やセグメント特徴量のように，長い時間範囲の情報を持った特徴量が一般に用いられている．長い時間範囲の情報が必要だというのに，分析を

†1 東北大学大学院工学研究科
Graduate School of Engineering, Tohoku University

なぜ短い時間単位で行う必要があるのだろうか．分析窓自体をもっと長くすれば，一つのフレームにより多くの情報を含めることが可能ではないのだろうか．その場合には，10ms よりももっとフレームシフトを長くすることが可能かもしれない．

以上の二つの疑問点について，本稿では小規模な音声認識実験を通して考察する．

2. フレームに基づく音声分析と音声認識

フレームシフトの値はなぜ 10ms 程度なのだろうか．最初期の音声認識の研究を調査すると，初期には 10ms ごとに零交差数を計算して特徴量とする方法[2,3]や，同じく零交差数を 1ms 程度の間隔で計算するもの[4]，アナログバンドパスフィルタの出力を 10ms ごとにサンプリングして特徴量とする研究[5]などが行われていた．特徴量抽出をすべて計算機によって行うようになっても，特徴量の時間間隔はおよそ 10ms である[6,7]．これらの研究では，特徴量の時間間隔がなぜ 10ms 程度なのかについての理由は特に述べられていないが，その理由はおそらく音素長分布に関係すると考えられる．アメリカ英語の子音の継続長を調べた Umeda の研究によれば[8]，子音の継続長は音素の単語内位置やストレスの有無に大きく影響されるが，短いものでおおむね 20ms 程度である．そのため，フレームシフトが 20ms を超える場合，1 フレーム内に複数の音素の情報が混在してしまい，認識が困難になると考えたと思われる[1]．より短いフレームシフトによる音声分析を行えば精密な分析が可能だと思われるが，フレームシフトが短ければ特徴量系列が長くなり，認識にかかる計算コストが増大する．そのため，計算コストとの兼ね合いを考えて 10ms 程度に落ち着いたと考えられる．

上記のとおり，フレームシフトが長ければ全体のフレーム数が少なくなるため，計算コストが下がる．そのため，計算コスト削減法としてフレームを間引くことがある．McLaulin らは，GMM を用いた話者認識タスクにおいて，10ms のフレームシフトで分析した特徴量を 1/4 程度に間引いても性能には影響がないとしている[9]．同様の処理は計算量削減法として主に話者認識・話者照合で使われている[10]．話者認識においてフレームの間引きが使われている理由は，話者認識の目標が音声スペクトルの全体的な分布の把握であるため，個々の音素を正確に把握できなくても大きな問題にならないためであろう．これらの方法では，10ms フレームシフトでモデル学習を行ったうえで，認識時に特徴量を間引くという処理を行っているので，フレームシフト自体を 20ms や 40ms にしているわけではない．

10ms よりも長い音声信号の影響を考えて音素をモデル化する方法も検討されており[11]，その 1 つである Δ 特徴量[12]は広く用いられている標準的な手法である．また，時間的に連続する特徴ベクトルをまとめて一つの特徴量とする方法（セグメント特徴量）も古くから検討されており[13]，3 層ニューラルネットを用いた最初の音声認識モデルである Time Delay Neural Network (TDNN)をはじめ[14]，最近の Deep Neural Network (DNN)を用いる音声認識[15]では標準的に用いられている．しかし，音声の分析窓を長くしたり，フレームシフト自体を長くする手法は，筆者の知る限り提案されていない．

3. 分析窓位置の偶然性の補償

3.1 分析窓のずれの影響

前述のとおり，通常の音声分析では，入力音声の 1 サンプル目を起点として分析窓をかけ，そのあとはフレームシフトの時間だけ分析窓をずらしながら分析を行う．このような分析においては，音声信号の時間変化に比べてフレームシフトが十分短く，分析窓のわずかなずれによってスペクトルに急激な変化が起きないということが仮定されていると考えられる．しかし，これは実際の音声信号で十分確認された事実とは考えにくい．図 1 は，16kHz でサンプリングされた「印鑑証明」という音声の/ka/の部分の音声波形を示している．赤い線は 25ms の分析窓を 10ms ずつずらして適用した場合の様子を示し，青い線は赤い線から 5ms ずらした場合の分析窓位置を示す．破裂子音は時間的な変動が大きいため，5ms のずれに対しても波形が比較的大きく変化することがわかる．

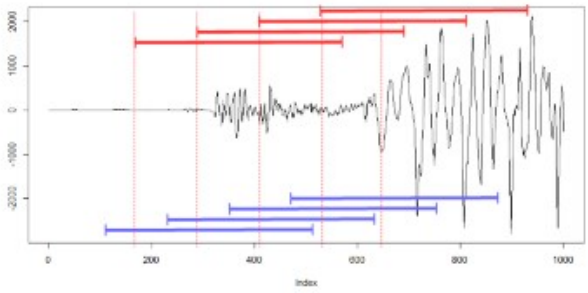

図 1　Waveform of /ka/ and analysis windows

同じ音声について，オリジナルの音声と，先頭の無音サンプルを削除した音声をそれぞれ MFCC 系列に変換し，その間で DP マッチングによって距離を計算した．分析窓は 25ms，フレームシフトは 10ms (160 サンプル)である．削除したサンプル数と DP 距離との関係を調べたものを図 2 に示す．分析対象は同じ音声であり，削除した先頭のサンプルは無音区間であるから，DP 距離は変化しないことが期待される．しかし実際には，図のように距離は比較的大きく変化し，その周期はフレームシフトに一致する．

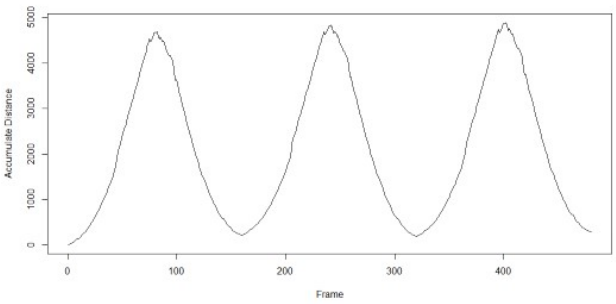

図 2 DP distance between the word /iNkaNsho:me:/ and the shifted signal

図 3 は，元音声と 5ms（80 サンプル）ずれた音声および 10ms（160 サンプル）ずれた音声とを DP マッチングによって対応させた場合の，フレームごとの特徴量間の二乗距離を示している．黒が 5ms，赤が 10ms ずれた場合の結果を示す．5ms ずれた音声については，所々に大きな距離のピークがみられる．これらのピークは音声の非定常部で観測される．/k/の破裂部分で距離が最大であり，距離の値は約 556 であった．同じ音声の/e:/と/N/の間の距離の最小値が約 370 であったから，分析窓のずれによって，全く同じ音声中での異なる音素間の距離に匹敵する違いが表れていることがわかる．

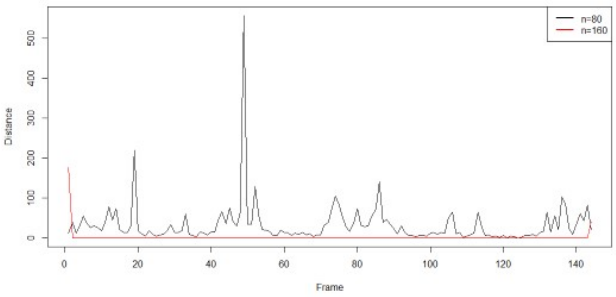

図 4 Frame-by-frame square distances between the same speech with different analysis window positions

3.2 分析窓の補償

通常の音声においては，同一の音素の特徴量には，ある話者の中での発声の変動や音素環境による変動，異なる話者の間での発声変動に加えて，分析窓の位置のずれによるスペクトル変動も含まれる．そのため，統計的な音響モデルの学習において，学習データ量が十分多ければ，分析窓のずれによる変動も含めて学習されると考えられる．しかし，学習データが十分でなかった場合には，分析窓の位置による変動がうまくモデル化されない可能性がある．

そこで，学習データ量が少ない場合に，分析窓の位置を様々にずらした音声を与えることにより，分析窓ずれによる特徴量の変動をモデル化することができると考えられる．音声のサンプリング周波数が 16kHz であった場合，フレームシフト 10ms は 160 サンプルに相当する．そこで，学習用の音声の先頭が無音であることを仮定した場合，先頭の 80 サンプルを削除して分析すると，分析窓の位置を 5ms ずらした場合の音声を作成することができる．同様に，分析窓の位置をずらした音声を複数作成し，これらを独立した学習データとして音響モデルの学習に用いる．

3.3 実験

実験用の学習・評価データとして，東北大・松下単語音声データベース [16]に含まれる単語発声（60 名×212 単語）を用いた．これを 30 名からなる 2 つのセットに分け，片方を学習用，もう片方を評価用として 2 分割交差検定を行った．単語セットは 212 単語で固定であるため，今回の実験は話者オープン・語彙クローズドの実験である．

音声は 16kHz サンプリング・16bit 量子化であり，これを 25ms のハミング窓で分析した．フレームシフトは 10ms である．特徴量は MFCC12 次元と対数パワー，および 1 次と 2 次の回帰係数（計 39 次元）である．

使用した音響モデルは GMM-HMM による triphone であり，認識タスクは音素認識である．認識に際しては，triphone の制約を満たす音素系列のみを認識するような言語モデルを用いた．評価においては，triphone の中心音素の正誤のみを判定に用いた．

学習データの作成においては，表 1 に示す 5 条件を比較した．例えば “Shift 4” の条件では，サンプルずれが 0, 40, 80,120 サンプルである 4 つのデータをそれぞれ学習データとして利用した．このため，この条件では学習データ量は 4 倍になる．

表 1　Conditions for sample shifts for training data preparation

条件	サンプルずれ（サンプル数）
Shift 1	0
Shift 2	0, 80
Shift 4	0, 40, 80, 120
Shift 8	0, 20, 40, 60, 80, 100, 120, 140
Shift 16	0, 10, 20, 30, 40, 50, 60, 70, 80, …, 150

実験結果を図 5 に示す．GMM の混合分布数を 1 から 8 まで変化させて実験を行った．全体として混合数 4 のときが良い性能であり，40 サンプルずらして学習データに加えた場合が最も性能が高く，音素正解精度は 92.8%であった．また，いずれの混合分布数においても，サンプルをずらしたデータを学習に加えることで認識精度が向上した．しかし，混合数 1 の場合はその影響は少なかった．

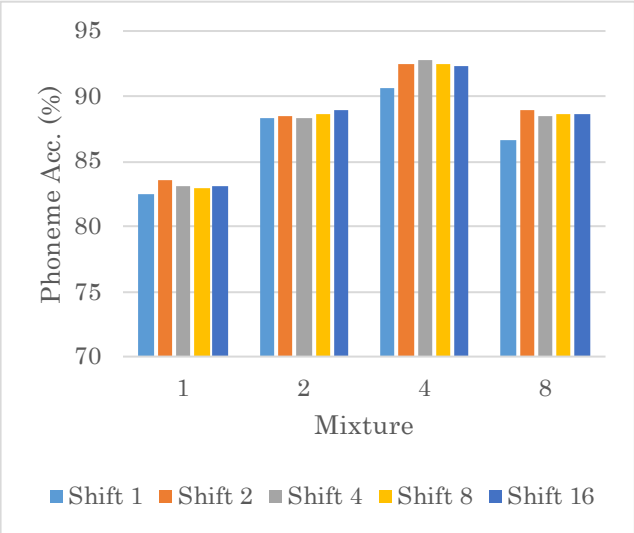

図 5 Recognition results for sample shift

4. 長い分析窓とフレームシフト

4.1 フレームシフトを長くすることのメリット

通常，フレームシフトは 10ms 前後（ 5ms, 8ms, 15ms など）であるが，前節の実験から分析窓の位置ずれによる影響が学習によって補償できるのであれば，もっとフレームシフトが長くても性能を下げずに認識できる可能性がある．フレームシフトが長ければ，認識のためのフレーム数が少なくなり，認識のための計算コストが小さくなるため，実際の音声認識システムにおいてはメリットがある．

4.2 実験条件

実験においては，分析窓長・フレームシフト・HMM の状態数を様々に変えて実験を行った．この時の条件を表 2 にまとめた．また，あるフレームシフト条件では，前節と同じく，“Shift 1”（通常の音声分析）～“Shift 16”（フレームシフトの 1/16 ずつずらした音声を学習に使用する）までの 5 条件を試した．フレームシフトが 40ms より長い条件においては，1 音素のフレーム数が 3 フレーム未満になるものが多く，音素 HMM を 3 状態とした場合には学習ができなかった．そのため，HMM の状態数を 2 と 1 に変えて実験を行った．また，各音素 1 状態にした場合の性能を比較するため，各条件において HMM を 1 状態としたときの性能も調べた．

4.3 実験結果

実験結果を図 6 に示す．結果が多いため，ここでは各条件の中で最も高い性能を示した GMM 混合数の結果のみを示している．横軸のラベルの中のカッコの中の数字（“25-10-3 (4)” の “4”など）は GMM 混合数である．この図において，25-10-3 の結果は図 5 の結果と同一である．グラフには 25-10-1 の結果が含まれていないが，この条件では音素の挿入が極めて多く，認識結果が悪かった（8 状態での音素正解率が最高 7.5%）ため，グラフからは除外している．

この結果から，サンプルをシフトさせたものを学習に用いることに効果があることがわかる．50-30-3 においては，学習データの増加によって音素正解精度が 73.5%から 89.9%まで改善する．性能が最も高かったのは，100-40-2 条件での 1 混合 HMM を使った場合であった．1 状態 HMM であっても，100-50-1 の条件では最高で 89.8%の音素正解精度が得られた．これは，通常の条件(25-10-3)でのサンプルずれ補償を行わなかった場合の最高性能 90.6%に迫る性能である．もし 100-50-1 の条件で音声認識が可能ならば，フレーム数が 1/5，HMM 状態数が 1/3 であり，大幅な計算時間の削減が期待される．実際の認識計算にはビーム幅などの要因が大きく影響するので，計算時間削減効果については今後さらに詳細な評価が必要である．

表 2 Experimental conditions for speech analysis

条件	分析窓(ms)	フレームシフト(ms)	状態数
25-10-3	25	10	3
50-20-3	50	20	3
50-30-3	50	30	3
100-40-2	100	40	2
100-50-2	100	50	2
25-10-1	25	10	1
50-20-1	50	20	1
100-40-1	100	40	1
100-50-1	100	50	1
100-60-1	100	60	1

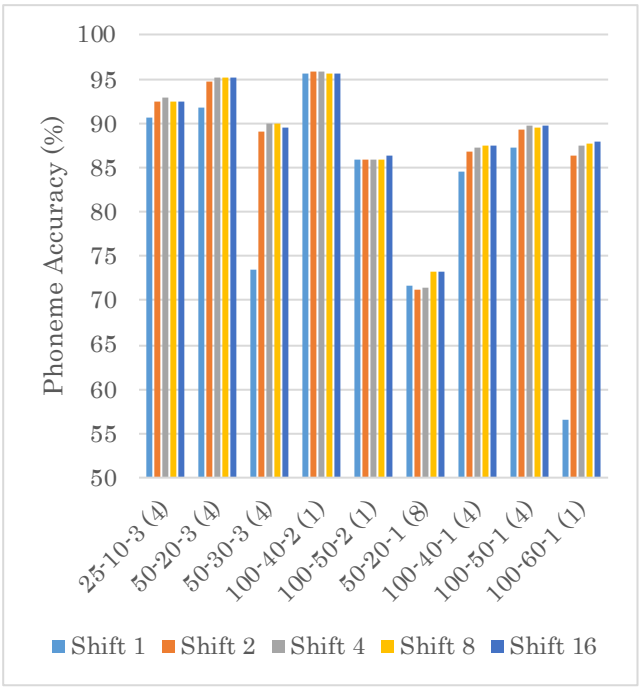

図 6 Phoneme recognition results for various window length,

frame shift and HMM state per phoneme

性能が高かった 100-40-2 の条件について，詳しい結果を図 7 に示す．図 5 に示したように，通常は GMM の混合数を増加させることで認識性能が向上するが，この結果では混合数を増加させても性能が向上しない．4 混合の Shift 1 条件，および 8 混合条件での認識性能が極端に低いが，この条件では学習ができない音素があり，それが原因で性能が下がっていることが確認できた．2 混合以上ではサンプルシフトによる学習には効果があるが，1 混合ではほとんど効果がない．これは図 5 の結果とも整合する．

同じく，100-50-1 の条件についての結果を図 8 に示す．この結果は図 7 の結果に似ていて，1 混合の性能が高く，この場合にはサンプルシフトの効果がない．4 混合の場合にはサンプルシフトの効果があり，Shift 4 の条件で認識性能が最大になる．混合数 8 では学習がうまくいっておらず，性能は低い．混合数 2 ではサンプルシフトによって性能が下がっているが，原因は不明である．

4.4 認識結果の観察

長いフレームシフト・少ない状態数で認識した結果がどうなっているのか，もう少し詳しく観察した．話者 609 の発声した単語/kurotabi/について，フレームシフト 20ms で性能の高かった 50-25-3 と，長いフレームシフトで性能の高かった 100-40-1 での認識結果（いずれも全音素正解）の音素境界を比較した．結果を図 9 に示す．図中の赤枠はデータベース付属のラベルによる音素位置を表す．この結果から，各 triphone は本来の音素境界を越えて，両端の音素のわたり部も含む形でモデル化されているらしいことがわかる．ただし，100-40-2 による k-u+r, u-r+o, r-o+t の境界を見ると，必ずしも中心音素をモデル化するように学習されていない．今回の実験は語彙クローズドであり，話速にも大きな変動がないので，このようにずれたセグメントを「音素」として学習してしまっても矛盾はないが，CSJ のように話速の変化が大きくて音素環境がオープンな（すなわち，未知の triphone を類似音素環境の triphone で代替する必要がある）条件において，長いフレームシフトによるモデル化が可能なのかどうかは今後の検討を必要とするであろう．

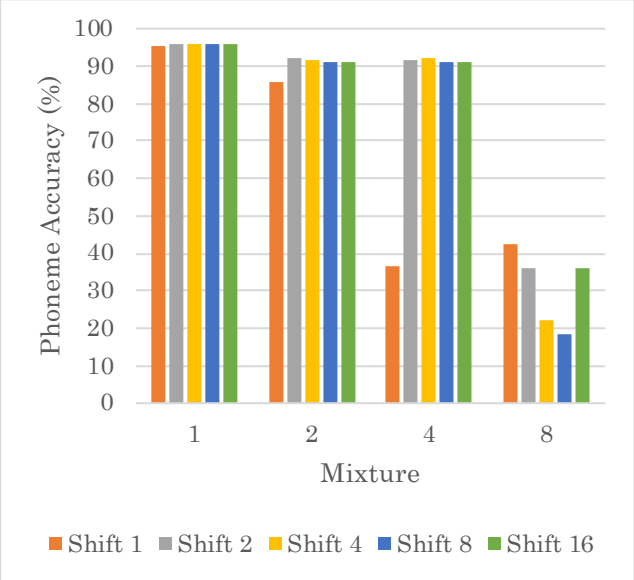

図 7 Recognition results for condition 100-40-2

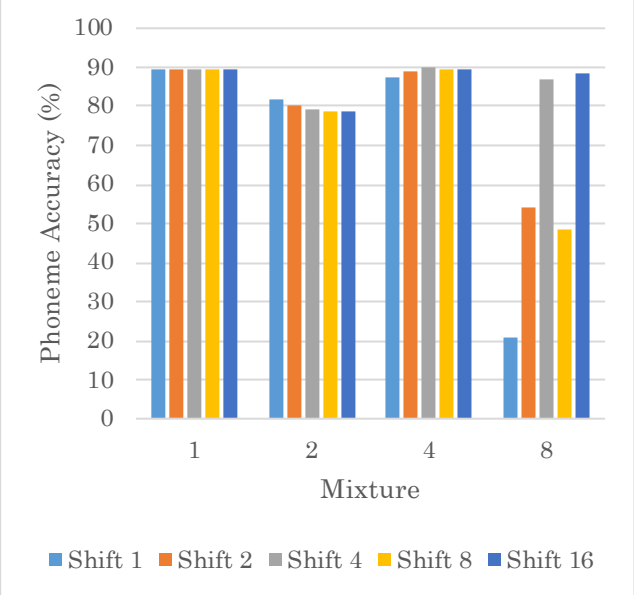

図 8 Recognition results for condition 100-50-1

5. まとめ

これまでほとんどの音声認識研究において用いられてきた「分析窓幅 25ms，フレームシフト 10ms」の条件について再考した．まず，フレームシフトが 10ms の場合，フレーム開始位置の変動によって，抽出される特徴量が大きく変動する場合があることを実験的に示し，これに基づいて，フレーム位置をずらした音声を学習データに加えることで認識性能を向上させる方法を提案した．次に，フレーム位置の変動を学習によって補償できることを前提に，フレームシフトを長くする実験を行った．実験の結果，フレームシフト 40ms の場合には 10ms の場合を超える性能が得られ，フレームシフト 50ms で 1 状態 HMM を用いても通常

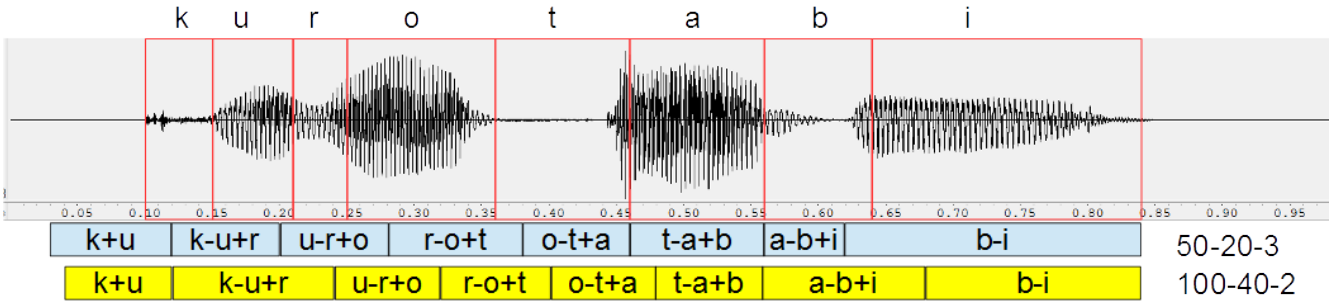

図 9 Alignment of triphones for word /kurotabi/ using 50-20-3 and 100-40-2 models

の認識条件に匹敵する性能が得られた．

今回の実験は小規模な単語音声による評価であり，しかも評価が音素環境クローズドであるため，今回の結果には実験条件に起因するアーティファクトが含まれている可能性がある．また，今回のフレーム位置の補償については，前述の通り学習データが大規模になればその効果が小さくなるだろうと予想される．これらの点について，今後評価を行っていきたいと考えている．

参考文献

1) O'Shaunessy, D.: Acoustic analysis of automatic speech recognition, Proceedings of IEEE, Vol. 101, No. 5, pp. 1038-1053, 2013.
2) Sakai, T. and Doshita, S.: The automatic speech recognition system for conversational sound, IEEE Trans. Electronic Computers, Vol. EC-12, No. 6, pp. 835-846, 1963.
3) Reddy, D. R.: Phoneme grouping in speech recognition, J. Acoust. Soc. Am., Vol. 41, No. 5, pp. 1295-1300, 1967.
4) Bezdel, W. and Bridle, J. S.: Speech recognition using zero-crossing measurements and sequence information, Proc. IEE, Vol. 116, No. 4, pp. 617-623, 1969.
5) Reddy, D. R., Erman, L. D. and Neely, R. B.: A model and a system for machine recognition of speech, IEEE Trans. Audio and Electroacoustics, Vol. 21, No. 3, pp. 229-231, 1973.
6) Itakura, F.: Minimum prediction residual principle applied to speech recognition, IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 23, No. 1, pp. 67-72, 1975.
7) Jelinek, F.: Continuous speech recognition by statistical methods, Proceedings of the IEEE, Vol. 64, No. 4, pp. 532-556,1976.
8) Umeda, N.: Consonant duration in American English, J. of Acoust. Soc. Am., Vol. 61, No. 3, pp. 846-858, 1977.
9) McLaulin, J., Reynolds, D. A. and Gleason, T.: A study of computation speed-ups of the GMM-UBM speaker recognition system, In: Proceedings of Eurospeech, pp. 1215-1218, 1999.
10) Kinnunen, T., Karpov, E. and Franti, P.: Real-time speaker identification and verification, IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No. 1, pp. 277-288, 2006.
11) Hermansky, H.: Exploring temporal domain for robustness in speech recognition. In: Proceedings of the 15th Int. Congress on Acoustics, Vol. II, pp. 61-64, 1995.
12) Furui, S.: Cepstral analysis technique for automatic speech verification, IEEE Trans. Acoustics, Speech and Signal Processing, pp. 254-272, 1981.
13) Makino, S., Kawabata, T. and Kido, K.: Recognition of consonant based on the perceptron model, In: Proceedings of Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pp. 738-741, 1983.
14) Weibel, A., Hanazawa, T., Hinton, G., Shikano, K. and Lang, K. J.: Phoneme recognition using time-delay neural networks, IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 37, No. 3, pp. 328-339, 1989.
15) Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. and Kingsbury, B.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, IEEE Signal Processing Magazine, Vol. 29, No. 6, pp. 82-97, 2012.
16) 牧野正三，二矢田勝行，真船裕雄，城戸健一："東北大ー松下単語音声データベース"，日本音響学会誌 Vol. 48, No. 12, pp. 899-905, 1992.