同時通訳や音声対話の実用化に向けた大語彙音声認識技術

Large-Vocabulary Speech Recognition Technologies for Achievement of Simultaneous Translation and Speech Dialog Systems

益子　貴史　　芦川　将之
■MASUKO Takashi　■ASHIKAWA Masayuki

音声翻訳や音声対話などを実用化するためには，重要な入力手段の一つである音声認識を様々な話題に対応できるよう大語彙化する必要がある。しかし，日々生み出されている大量の新語や造語，口語表現などの語彙を，少数の開発者が収集し追加することは困難である。また，語彙追加に伴う類似単語の増加に対応するため，音韻識別性能の向上が不可欠である。

東芝は，これらの問題を解決するため，クラウドソーシングによる語彙収集方法を確立するとともに，音韻識別性能を向上させるための新たな音響特徴量を開発した。これにより，短期間に大量の語彙を収集できるようにするとともに音声認識精度を向上させ，大語彙音声認識を実用化した。

In order to achieve the practical use of voice translation and speech dialog systems, large-vocabulary speech recognition that recognizes utterances of various types is required. However, it is difficult for a small number of developers to collect the new words and colloquial expressions that continuously appear in the language and to add them to a system's vocabulary. Moreover, it is necessary to improve phoneme discrimination performance in order to discriminate between the increasing number of similarly pronounced words that emerge with the expansion of vocabulary size.

To overcome these problems, Toshiba has established a word collection method using crowdsourcing and developed a new acoustic feature to improve phoneme discrimination ability. These technologies realize large-vocabulary speech recognition through the collection of a number of words in a short period of time and improved speech recognition accuracy.

1　まえがき

東芝はこれまで，カーナビなどを音声で操作するためのコマンド音声認識エンジンや，スマートフォン向け音声翻訳システム[(1)]のための連続音声認識エンジンを開発してきた。これらの音声認識エンジンでは，認識対象を少数のコマンド，あるいは旅行会話などの特定の内容に限定することにより，少ない計算量で高い認識性能が得られている。

一方，現在開発中の同時通訳システム（この特集のp.18－21参照）や音声対話システム（同p.14－17参照）では，あらかじめ用途を限定していたとしても，実際には雑談など想定用途外の発話が入力される場合も多い。そのため，入力音声を文字化する音声認識エンジンも幅広い話題の発話を高い精度で認識できる必要がある。そこで当社は，近年音声認識エンジンの大語彙化(注1)と高精度化に取り組んでいる。

2　大語彙化と高精度化への取組み

音声認識エンジンの構成，及びこれまで行ってきた各構成要素に対する主な性能改善アプローチを**図1**に示す。これらのアプローチを代表として様々な性能改善を行うことにより，大語彙音声認識でも高い認識性能が得られるようになってきた。

音声認識エンジンの大語彙化のためには，幅広い話題に対するカバー率が高くなるように，語彙を単語辞書から選択し単語モデルに追加する必要がある。更に，新たな話題に対応す

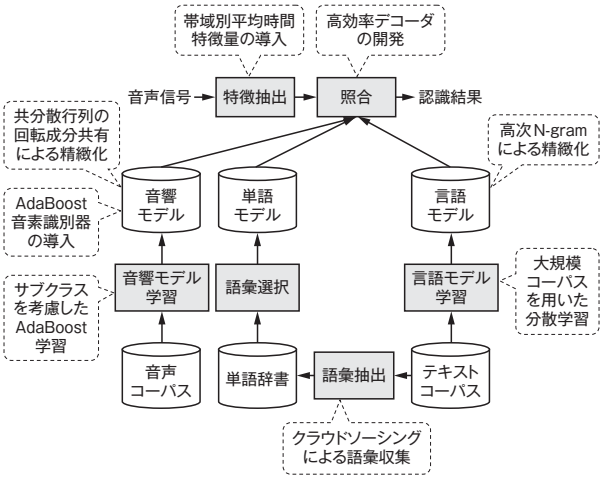

図1．音声認識エンジンの構成と性能改善アプローチ ― 音声認識エンジンの構成要素に対して，これまで様々な性能改善手法を検討してきた。
Configuration of speech recognition engine and approaches to performance improvement

（注1）　認識可能な単語の種類を増加すること。

るためには，日々生み出される新語や造語，口語表現などの大量の語彙を継続して収集し，単語辞書及び単語モデルに追加しなければならない。しかし，少数の開発者がそのような大量の語彙を収集することは現実には困難である。そこで，多数の一般人が作業を行うクラウドソーシングを用いた語彙の収集手法を開発した[2]。

また，認識語彙の増加に伴って発音の類似した語彙も増加することから，十分な認識精度を得るためには，音韻識別と単語予測の性能向上が必要になる。単語予測については，大規模なテキストコーパス(注2)を用いた高次N-gram言語モデルの分散学習により性能向上を図っている。一方，音韻識別性能の向上については，共分散行列の回転成分の共有化による音響モデルの精緻化[3]，AdaBoost音素識別器を用いたリスコアリング手法の導入[4]，及びサブクラスを考慮したAdaBoost音素識別器の学習手法の開発[5]に加え，新たな音響特徴量である帯域別平均時間ケプストラム（SATC：Sub-band Average Time Cepstrum）を開発した[6]。

これらの性能改善アプローチのうち，クラウドソーシングによる語彙収集とSATCについて，以下に概要を述べる。

3　クラウドソーシングによる語彙収集

3.1　クラウドソーシングとその課題

従来，テキストコーパスからの未知語の自動獲得手法が研究されているものの，例えば読み推定精度は90％程度である[7]など，自動獲得だけでは，得られた語彙をそのまま辞書として利用できるレベルには至っていない。そのため，辞書として利用できる精度の語彙を獲得するためには，人手によるチェックが不可欠である。しかし，これを少数の開発者で行うことは現実的ではない。これを解決する手段として，クラウドソーシングの活用を検討した[2]。

クラウドソーシングとは，単純ではあるが自動化することが困難な作業（タスク）を，不特定多数の一般の作業者に委託し実施してもらう手法である。このクラウドソーシングにより，少数の専門家や開発者だけでは実施が困難な大量のタスクを短期間で行うことが可能になる。

既存クラウドソーシングサービスの問題の一つとして，作業結果の精度が低いという点が挙げられる。クラウドソーシングでは，専門家ではない一般作業者にタスクを依頼するため，もともと専門家に比べて精度が低いという問題がある。更に，タスクをいいかげんに実施する不誠実な作業者も少なからず存在し，それが全体的な精度低下の大きな要因となる。そのため，いかに個々の作業者の作業結果の精度を高めるかが重要な課題となる。

この問題を解決するため，これまで当社が開発してきたPrivate CrowdSourcing System（PCSS）[8]を語彙収集に利用した。

3.2　PCSS

PCSSでは，作業者を募集する際に，学歴や1日当たりの作業可能時間などのユーザーの属性に基づいて，必要な条件に合致するユーザーをあらかじめ抽出することで，ある程度作業結果の精度を向上させることができる。しかし，クラウドソーシングの作業内容は多岐にわたるため，事前の調査だけでは十分ではない。そこで，作業者がPCSSで作業を開始してからの行動履歴をベースに，作業者の正解率と経験値，及び“スキル”の管理による精度向上を試みている。

正解率は“正解数/作業数”で算出し，一定値以下の作業者は作業不可にすることで作業結果の精度を向上させる。また同様に“正解数−不正解数”で算出される経験値を設定している。一定の経験値を持つユーザーに対して高報酬で高難易度の作業を提供することで，好成績のユーザーのモチベーションを高める目的がある。

スキルとは，作業者の作業結果から判明した，“文法に詳しい”や“音感が良い”などの（作業者自身が自覚していない場合がある）特徴である。例えば“読みがな付け”の作業の正解率が高い作業者には読みがな付けのスキルを付与し，以後の読みがな付けの作業を読みがな付けスキルを持つ作業者に優先して出題することで精度を向上させる。

3.3　PCSSによる語彙収集

PCSSを用いた語彙収集では，まずWebクローリング(注3)で収集した大量のテキストから，テキスト処理により未知語候補を抽出する。この未知語候補には単語として適当でないものが含まれている可能性が高い。また抽出した単語に対して音声処理に必要な情報を付与しなくてはならない。そこでPCSSを用いて，未知語候補に対する単語判定，品詞付与，及び読み付与の各タスクを順次行った。

単語判定では，未知語候補を“それは（未知語候補）です”という問題文に加工して表示し，“問題文は日本語として自然か否か”という選択をさせた。そして，“自然である”と回答された文に含まれる未知語候補を未知語として扱った。品詞付与では，名詞とそれ以外の品詞に分ける作業を行い，名詞に関しては更に“人名”，“地名”，“組織名”，及び“その他の名詞”に再分類した。読み付与では，名詞と判定された未知語に対して，その読みを入力させた。各タスクは3人に出題され，2人以上一致した回答を有効なデータとして扱った。ただし，単語判定は高精度であることを求められるため，3人が一致した回答だけを有効なデータとして扱った。

実際にPCSSによる語彙収集を実施したところ，125億文の

（注2）　自然言語の文章を大規模に収集し，言語的な情報を付与したもの。

（注3）　Web上を自動的に巡回してWebページを収集すること。

表1．作業結果の一致率

Concordance rates of vocabulary collection activities

タスク	3人一致	2人一致	不一致	未回答*
品詞付与（%）	71.3	27.7	0.9	0.1
読み付与（%）	82.8	11.2	1.5	4.5

＊1～3人が未回答で，3人分の回答がそろわなかったもの

Webテキストから抽出された約23万語の未知語候補から，約14万語を新たな語彙として収集できた。品詞付与及び読み付与における作業結果の一致率を，**表1**に示す。2人以上が一致したデータからランダムに約1,000語を抽出して正解精度の評価を行ったところ，品詞付与と読み付与ともに，98～99％の非常に高い正解率を得ることができた。実際に獲得できた未知語の例としては，“Siri”や，“あっちゃん”，“先っちょ”，“スンゲー”などが挙げられる。

4　音韻識別性能向上のための新たな音響特徴量

4.1　新特徴量の開発方針

認識語彙を増加させると，発音の類似した語彙も増加する。そのため，十分な音声認識精度を得るためには，言語モデルの性能向上とともに，類似単語を識別するための音韻識別性能の向上が必要になる。そこで，当社は新たな特徴量の開発を試みた。

新特徴量を開発するにあたり，新特徴量は従来特徴量と相補的な性質を持つものであることを指針とした。MFCC（Mel-Frequency Cepstral Coefficient）などの従来特徴量の多くは，短時間分析窓を用いたスペクトル分析によって求められる振幅スペクトル情報に基づく特徴量である。そのため，従来特徴量では長時間にわたる音声の情報を十分に表現することができない。これに対し，新特徴量は従来よりも長い時間の分析窓を用いて求められ，振幅スペクトル以外の情報に基づく特徴量とした。これにより，従来特徴量では捉えられていない情報を扱えるようになり，従来特徴量と併用することにより音韻識別性能の向上が期待できる。

4.2　SATC

そのような特徴量として，当社はSATCを開発した[6]。平均時間とは信号の時間軸上での重心位置（単位は時間）である。帯域別平均時間は，従来よりも長い分析窓で切り出した信号の，周波数帯域ごとの時間軸上での重心を求めたものである。SATCは，この帯域別平均時間にDCT（Discrete Cosine Transform）を適用して得られる特徴量である。

従来の多くの特徴量が基づいている振幅スペクトル包絡と，帯域別平均時間の概念を**図2**に示す。(a)は音声信号の時間波形と，周波数分析を行って得られた時間周波数平面を表している。時間周波数平面の縦軸は周波数，横軸は時間で，青色は振幅が小さく，赤色は振幅が大きいことを表している。また，(b)は短時間分析窓で切り出された信号の振幅スペクトルとその包絡を示している。縦軸は周波数，横軸は振幅，黒色の破線は周波数帯域の境界，黒色の実線は振幅スペクトル，赤色の×印は周波数帯域ごとの平均振幅，赤色の実線は振幅スペクトルの包絡を表している。ただし，振幅は対数をとったものである。一方，(c)は長時間分析窓で切り出された信号の時間周波数平面と帯域別平均時間を示している。縦軸は周波数，横軸は時間，黒色の破線は周波数帯域の境界，赤色の×印は周波数帯域ごとの平均時間，赤色の実線は帯域別平均時間の包絡を示している。

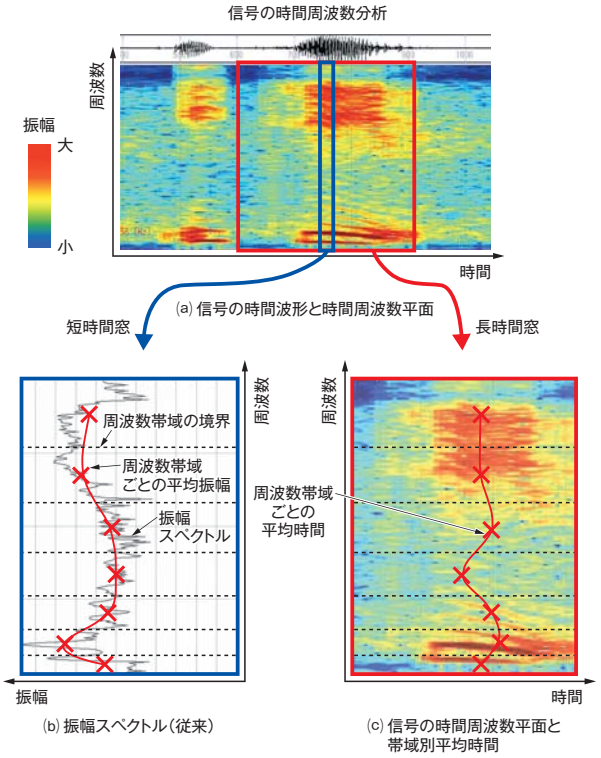

図2．振幅スペクトル包絡と帯域別平均時間 ― 短時間窓で求められる振幅スペクトルに基づく従来特徴量と，長時間窓で求められる帯域別平均時間に基づくSATCを併用することで，音韻識別性能を向上させる。

Amplitude spectrum envelope and sub-band average time

振幅スペクトル包絡と帯域別平均時間の包絡は，ともに縦軸（次元）は周波数帯域で共通しているものの，横軸（値）はそれぞれ振幅と時間で互いに異なる情報を表現しており，包絡の形状も異なっている。

平均振幅と帯域別平均時間との関係を示すため，ある帯域での平均振幅と帯域別平均時間の時間変化のようすを**図3**に示す。横軸は時間を表し，縦軸は上段が平均振幅，下段が帯域別平均時間を表している。帯域別平均時間は，一点鎖線で示している平均振幅のピークを中心に，時間が進むに従って値が減少していることがわかる。これは，分析窓が音声波形

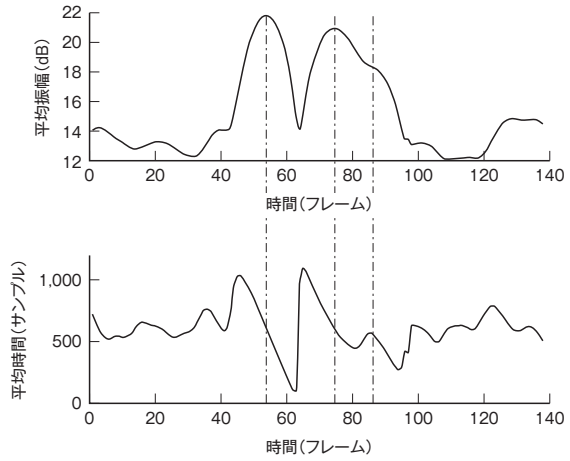

図3. 平均振幅と帯域別平均時間の関係 — 平均振幅のピークを中心に平均時間の顕著な減少が見られる。
Relationship between amplitude and sub-band average time

上で過去から未来に進むにつれて，分析窓内で振幅のピーク部分が未来から過去に移動していくためである。

4.3 新特徴量の評価

この新たな特徴量SATCを大語彙連続音声認識において評価した。評価データは，Webから収集した最新のニュース及びブログの読上げ音声と，コンタクトセンター（CC）のオペレーターによる対話音声である。従来特徴量であるMFCCだけを用いた場合とMFCCとSATCを併用した場合の文字正解精度（CAcc）及び誤り削減率（RERR）を**表2**に示す。

SATCの併用による誤り削減率は，Web読上げ音声に対しては7％，CCの対話音声に対しては18％となった。読上げ音声は滑舌よく発声されており，音響的な識別が比較的容易なため，従来特徴量であるMFCCだけでも高い認識性能が得られている。一方，対話音声は読上げ音声よりも発音が不明瞭になりやすく，音響的な識別がより困難になるため，新特徴量であるSATCを併用することによる認識精度の改善効果が大きく表れている。

表2. 文字正解精度と誤り削減率
Character accuracy and relative error reduction rates

評価対象	CAcc (%)		RERR (%)
	MFCC	MFCC+SATC	
Webの読上げ音声	89.43	90.17	7.00
CCの対話音声	75.21	79.81	18.56

5 あとがき

同時通訳や音声翻訳を実用化するために必須となる音声認識の大語彙化のためのクラウドソーシングによる語彙収集方法と，大語彙音声認識の精度向上のための新たな特徴量であるSATCについて述べた。これらを含む様々な性能改善手法により，大語彙音声認識の性能向上を実現した。

今後も大語彙化と高精度化を進め，同時通訳や音声対話を含めた音声認識応用技術の実用化に貢献していく。

文 献

(1) 井阪岳彦 他．スマートフォン向け日中英音声翻訳システム．東芝レビュー．**65**，8，2010，p.48－51.
(2) 芦川将之 他．“PrivateCrowdSourcingを用いた言語，音声資源の収集”．人工知能学会全国大会（第27回）．富山，2013-06．3M3-OS-07d-2.
(3) Shinohara, Y. "Tying rotations of covariance matrices via Riemannian subspace clustering". Proc. ICASSP-2013. Vancouver, Canada, 2013-05. p.7000－7004.
(4) 藤村浩司 他．“AdaBoost音素識別器によるNベストリスコアリングの検討”．日本音響学会2011年春季研究発表会講演論文集．東京，2011-03．p.13－15.
(5) Fujimura, H. et al. "N-best rescoring by phoneme classifiers using subclass AdaBoost algorithm". Proc. INTERSPEECH 2013. Lyon, France, 2013-08, ISCA. p.3327－3331.
(6) 中村匡伸 他．“群遅延に基づく音声特徴量の雑音環境下での評価”．日本音響学会2012年春季研究発表会講演論文集．横浜，2012-03．p.135－136.
(7) 羽鳥 潤 他．“機械翻訳手法に基づいた日本語の読み推定”．言語処理学会第17回年次大会．豊橋，2011-03．p.579－582.
(8) 芦川将之 他．CrowdSourcingを用いた単語への読み付け，アクセント付け手法の提案．電子情報通信学会技術研究報告．AI，人工知能と知識処理．**111**，447，2012，p.11－16.

益子 貴史 MASUKO Takashi, D.Eng.
研究開発センター 知識メディアラボラトリー主任研究員，博士（工学）。音声情報処理の研究・開発に従事。電子情報通信学会，日本音響学会，IEEE，ISCA会員。
Knowledge Media Lab.

芦川 将之 ASHIKAWA Masayuki
研究開発センター 知識メディアラボラトリー研究主務。大規模データ処理の研究・開発に従事。人工知能学会会員。
Knowledge Media Lab.