3M3-OS-07d-3

PrivateCrowdSourcingを用いた言語、音声資源の収集 ～音声収集と品質評価～

A Devisable Private CrowdSourcing System for Speech Collection

中田 康太[*1] 芦川 将之[*1]
Kouta Nakata Masayuki Ashikawa

[*1]株式会社東芝研究開発センター
Corporate Research & Development Center, TOSHIBA

Collection of large speech corpora is often difficult due to its economical and temporal costs. We make use of a Private CrowdSourcing System and collect speeches from various people working on their own PCs. Since non-experts are prone to making errors, quality control is essential for crowdsourcing. We show that a post quality control keeps quality of speeches as high as carefully recorded materials for Automatic Speech Recognition. We also introduce online quality control by a "scoring system" and improve efficiency of recording.

1. はじめに

音声処理の学習や評価に利用する音声の収集は経済的・時間的コストが大きく、短時間に大量の音声を収集することが困難であることが知られている。本研究ではCrowdSourcingにより安価かつ迅速に音声を収集して音声処理に利用することで、音声処理の基本性能を向上することを目的としている。

本報告では、報告者らが構築したPrivate CrowdSourcing System ([Ashikawa 12]) に音声を収集する環境を整備し、実際に一般の作業者から音声を収集する。CrowdSourcingで得られた音声は必ずしも品質が高いとは限らないため、収集した音声を詳細に分析し、低品質の音声をフィルタリングする手法を確立する。フィルタリングにより得られた音声が音声認識エンジンの評価データとしての利用可能であることを示す。また音声収集作業において採点システムによるゲーミフィケーションを行い、音声の収集精度の向上を図る。

本報告の構成を以下に示す。2章ではCrowdSourcingを利用した音声収集の関連研究について報告する。3章では本研究で行った音声収集の概要について報告する。4章では収集された音声を分析し、品質低下要因である読み誤りを含む音声のフィルタリングについて報告する。5章ではフィルタリングの効果により、音声が音声認識エンジンの評価音声として利用可能であることを示す。6章では採点システムによるゲーミフィケーションを導入した実験により、収集精度が向上することを示す。7章で本報告書をまとめ、今後の課題について議論する。

2. 関連研究

CrowdSourcingを利用した音声収集についての報告は、2010年前後から増加している (e.g. [Parent 11])。CrowdSourcingで音声を収集する動機として、音声処理の強化と多言語展開の加速が挙げられる。CrowdSourcingでは大量の音声を安価で収集することが可能であるため、収集した音声を、音声認識エンジンの学習データ ([McGraw 09], [McGraw 11], [Freitas 10])、音声検索システムの評価データ ([Li 10])、音声対話システムの対話データ ([Parent 11])、発音訓練システム用音声データ ([McGraw 11]) に用いることで、音声処理の基本性能向上を図ることができる。またサービスの多言語展開を行う際には、リソースの少ない新たな言語についてCrowdSourcingにより迅速に音声を収集し、サービスの加速を図ることができる ([Ledlie 10])。

CrowdSourcingを利用した音声収集方法で最も一般的な方法は、テキストプロンプトを利用する方法である ([Lane 10])。テキストプロンプトを利用した例では、画面に作業者が読み上げるテキストが表示され、作業者は録音ボタンを押してテキストを読み上げることで音声を録音する。他の方法として、対話シナリオに沿って作業者に会話をさせる方法 ([McGraw 10])、手本となる音声を聞かせて作業者に繰り返させる方法 ([Ledlie 10])、作業者に写真を提示して描写を発話させる方法 ([McGraw 11]) が提案されている。またCrowdSourcingと関連してゲームを利用した音声収集も行われており、非母国語の学習ゲームによる音声収集 ([McGraw 09])、音声対話を行うオンラインゲームによる音声収集 ([Chernova 10])、クイズゲームや音声合成の個人適応サービスを利用した音声収集 ([Freitas 10]) が提案されている。

CrowdSourcingにより収集される音声の品質は、高いコストをかけて収集される音声と比較して必ずしも高くない。例えば読み上げ音声の収集では、作業者の読み誤りが大きな問題となる。文献 [Lane 10] ではテキストプロンプトによる収録において、習熟度が低い作業者の読み上げ音声のうち約35%に読み誤りを含むことが報告されており、音声処理に利用可能な音声を収録するためには品質管理が不可欠である。文献 [Parent 11] では29例の音声関連のCrowdSourcingについてサーベイがなされており、多くの場合でCrowdSourcingを再度利用した内容確認などの品質管理を行っていることが報告されている。

CrowdSourcingによる音声収集の動機は明確である一方で、実際にCrowdSourcingの音声を音声処理に応用した例は少ない。文献 [McGraw 09] は、約10時間の音声を開発データとして音響モデル適応に利用している。また文献 [McGraw 11] は、1000発話の音声を音声検索エンジンの評価データとして用いている。他の報告では収集した音声の正確さについて検証するにとどまっており、CrowdSourcingによる音声の利用方法は確立されていないと言える。

3. CrowdSourcingによる音声収集

本研究では、報告者らが独自に構築したCrowdSourcing環境を利用して音声を収集する。図1は、本研究で行った音声収

連絡先: 中田康太，kouta.nakata@toshiba.co.jp

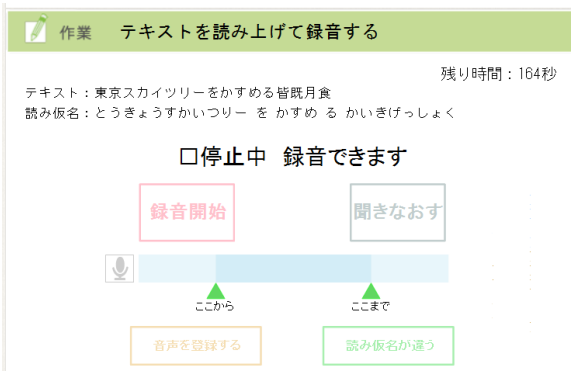

図 1: 音声収集作業画面の例

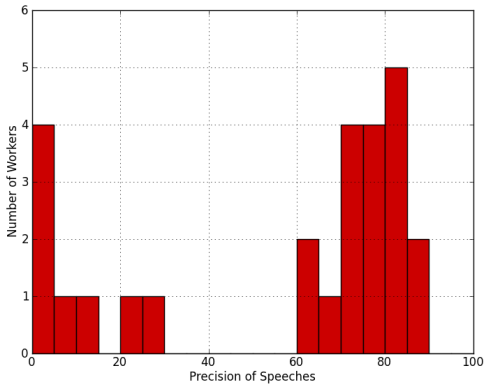

図 2: 作業者ごとの読み上げ正解率ヒストグラム

集作業の例を表している。音声の収録にはテキストプロンプトを利用し、作業者は録音ボタンを押してテキストの内容を読み上げることで録音を行う。録音された音声はインターネットを介してサーバに送信される。作業者は自分の音声を聞きなおすことも可能であり、読み間違えた場合は何度でも録音をしなおすことが可能である。作業の開始前にはマイク入力の値を用いた簡単な SN 比推定を行っており、ノート PC のオンボードマイクを用いている作業者や、録音環境が極端に悪い作業者を除外している。

本報告では、CrowdSourcing にニュースの読み上げ作業を出題することにより、29 名の作業者から 1812 発話の音声を収集した。読み上げの内容は 2011 年後半のニュースで、時事的な内容や固有名詞を含んだ内容になっている。報酬は一般的に高品質の音声を収録する一発話あたりのコストの 1/10 程度に設定した。収録にかかった期間は約 2 日間であった。

表 1 は、ニュース読み上げ作業を行った作業者の数を性別、年齢層別に表している。Private CrowdSourcing System には 20 歳代から 60 歳代の男女の作業者がいるが、今回の音声収録作業に関して 60 歳代の作業者はいなかった。作業者には女性が多く含まれており、特に 30 歳代の女性の作業者が多い。音声処理に向けて音声を収集する場合には性別、年代が均等になるように調整することが多いため、更に大量の音声を収集する際には性別と年代の偏りに注意する必要がある。

	20 歳代	30 歳代	40 歳代	50 歳代	60 歳代
男性	2	4	4	0	0
女性	1	10	4	4	0

表 1: 音声収録作業者の性別と年齢層

4. 読み誤り除去による品質改善

高いコストをかけて読み上げ音声を収録する場合には、雑音環境、録音機器、話者の作業状況などの条件を制御することが可能であり、目的に好適な環境で正確な高品質音声を収録することができる。CrowdSourcing ではこれらの条件を制御することは難しく、必ずしも品質の高い音声が得られるとは限らない。本報告では、特に作業者の読み誤りに注目して品質管理を行う。本章では適切なフィルタリングを設計して読み誤りを含む音声を除くことで、音声処理に利用可能な品質の音声を抽出する。なお録音時の録音環境については、作業開始前に SN 比により極端に環境の悪い作業者を除外しており、またある程度の雑音は音声処理を利用する実環境を反映していると考えられるため、本報告ではフィルタリングの対象としない。

図 1 の音声収集では、読み直しの判断は作業者に委ねられており、読みを誤った場合でも次の発話に進むことができる。そのため作業者は読み誤り音声をそのまま収録することがあり、音声の品質低下を引き起こしている。ここでは再度 CrowdSourcing を利用し、音声正誤判定作業をフィルターとして用いることで読み誤り音声を取り除く。

音声正誤判定作業では作業者に CrowdSourcing で収集した音声とテキストを提示し、テキストと音声が一致しているかを判定してもらう。音声ごとに 3 名の作業者が判定を行い、3 名全員が一致していると判定した場合は正確な音声、そうでない場合に読み誤り音声とする。音声収集の作業者のプライバシー保護のために、音声正誤判定の作業者に聞かせる音声の音質を変化させ、個人を特定できないようにしている。

図 2 は、作業者ごとの正解率のヒストグラムを表している。ここで横軸は正解率、縦軸は作業者の数である。正解率は作業者が収録した音声数に占める正確な音声数の割合である。正解率は 10 音声以上を録音した作業者について算出している。図 2 の作業者の平均の正解率は 70.9%で、収集した発話の約 30%が読み誤りであることが分かる。比較的正確に収録を行っている作業者でも正解率は 80% 程度となっており、一度も正しく発話できていない作業者も存在する。正解率の低い作業者については他の作業における実績から信頼できる作業者であることが確認されており、今回の収集では純粋な読み誤りにより品質が低下していると考えられる。

音声正誤判定により読み誤りを除去することで、正確に読み上げられた高品質な音声を得ることができる。一方で、大規模に音声を収集する場合には音声正誤判定にかかる経済的・時間的コストも大きくなると考えられる。本報告では、事前に強制アラインメント (Forced Alignment, FA) による自動正誤推定を行うことで、音声正誤判定のコスト削減を行う。

FA は音声認識技術を利用してテキストと音声を対応させることで、テキストの各音素が発声された時刻を出力する。読み誤りによりテキストと音声の内容が異なる場合には、FA による適切な対応結果が得られず、音素の継続長が不自然に長いまたは短い値になる傾向がある。そのため本実験では、テキストの音素数 n と FA により異常な継続長が推定された音素数 n_{out} の比率により、収録音声の自動正誤判定を行う。ここで

は音素の継続長 t が $0.05 \leq t \leq 0.2$ の時に正常な継続長、それ以外の場合を異常な継続長であるとする。異常継続長率 r_{out} を $r_{out} = n_{out}/n$ で定義し、ある発話の r_{out} が閾値 r_{th} 以上の場合には誤った発話、r_{out} が r_{th} 未満の場合には正確な発話と判定する。

		正解ラベル	
		正確	誤り
FA 推定	正確	1209	320
	誤り	1	292

表 2: FA による自動正誤判定の Confision Matrix ($r_{th} = 0.20$)

表 2 は、閾値 $r_{th} = 0.20$ の時の自動正誤判定の Confusion Matrix を表している。ここで行は FA の推定による発話の正誤、列は実際の正誤 (正解ラベル) を表している。閾値 r_{th} を高い値に設定することで再現率を優先し、全体の再現率は 99.9%、適合率は 79.0% となっている。FA により正確と判定された 1529 音声のうち、実際に正確な音声は 1209 音声含まれている。一方で、読み誤りと判定された 293 音声については、正確な音声が 1 音声含まれているものの、ほぼ全てが読み誤り音声になっている。読み誤りと自動判定された 320 音声をフィルタリングし、正確な音声と判定された 1529 音声を CrowdSourcing の音声正誤判定により判定することで、読み誤り音声を全てフィルタリングすることができる。この場合、CrowdSourcing で正誤判定する音声の数は従来は 1812 発話から 1529 発話に減少し、15.6% のコスト削減を行うことができる。

5. 音声認識の評価音声としての利用検討

2 章で述べたように、CrowdSourcing を利用した音声の収集は積極的に行われているものの、音声処理のタスクに応用して実用的な効果を確認した例は少ない。本章では CrowdSourcing で収集した音声を評価音声として利用することで、CrowdSourcing による音声収集の有用性を示す。

本実験では音声認識エンジンの言語モデルに注目する。近年は音声認識技術の実用化に向け、新たな語彙を含んだ言語モデルや利用目的に特化した言語モデルが継続的に更新されるケースが増えている。言語モデルが更新された際には音声認識エンジンの性能を評価する必要があり、その度に評価音声を収集するコストも大きくなっている。CrowdSourcing で収集した音声が評価音声として利用できれば、多数の話者から安価かつ迅速に評価音声を収集することが可能になり、音声認識エンジン開発の加速化につながると考えられる。

	CrowdSourcing		高品質
	フィルタなし	フィルタあり	-
発話数	1812	1209	1000
ASRv30k	33.41	25.20	23.49
ASRv100k	23.51	14.35	12.80
誤り改善率	29.63	43.06	45.51

表 3: 収集音声の誤認識率と誤り改善率

表 3 は、CrowdSourcing で収集した音声の誤認識率を表している。ここで誤認識率は読み誤り音声フィルタリングの有無のそれぞれの場合について算出する。また参考としてコストをかけて収集した高品質の音声についての誤認識率を算出する。高品質音声は認識に好適な環境で録音されており、標準的な評価音声として使用することが可能である。高品質音声の読み上げ内容は CrowdSourcing で収集した音声と同一である。

表 3 では、2 つの音声認識エンジン ASRv30K と ASRv100K の誤認識率を算出している。ASRv30k、ASRv100k は同一の音響モデルを用いており、認識語彙と言語モデルのみが異なる。ASRv30k は認識語彙が 3 万語で、言語モデルはウェブの記事約 4000 万文を用いて学習している。ASRv30k は認識語彙が 10 万語で、言語モデルはウェブ記事約 5 億文を用いて学習している。

CrowdSourcing の音声に対してフィルタリングを行わない場合には、ASRv30k, ASRv100k の両方で誤認識率が非常に高くなっている。これは、評価音声に読み誤り音声が混入していることで、読み誤り音声の認識結果が誤認識率に反映されてしまい、音声認識エンジンの正当な評価ができていないと考えられる。読み誤りフィルタリングを組み合わせることで誤認識は大きく改善し、高品質音声と同等の値となる。

表 3 に、ASRv30k から ASRv100k に更新した際の誤り改善率を示す。誤り改善率は 2 つの音声認識エンジンの相対的な改善を示すために用いられる値であり、今回の実験では ASRv30k の誤認識率 e_{30k} と ASRv100k の誤認識率 e_{100k} を用いて $(e_{30k} - e_{100k})/e_{30k}$ により算出される。読み誤り音声を除いた CrowdSourcing 音声の誤り改善率の値は高品質音声と同等になっており、エンジンの性能改善を適切に反映した評価が実行できていると考えられる。この結果から、CrowdSourcing で収集した音声を適切にフィルタリングすることで、評価音声として利用可能な高い品質の音声が得られることが分かる。

6. ゲーミフィケーションによる収集精度向上

4 章では、FA により音声の自動正誤判定を行うことで、CrowdSourcing の音声正誤判定のコスト削減が可能であることを示した。一方で、CrowdSourcing では音声正誤判定よりも音声収集の方が報酬が高く、FA による後処理のコスト削減の効果は比較的小さい。本実験では、音声収集の読み誤りが 30%以上と高い値になっており、収集の段階で読み誤りを削減することが更なるコスト削減につながると考えられる。

音声収集時に読み誤りを削減する効果的な方法として、オンラインで FA を利用する方法が考えられる。例えば、作業者が発話した直後に FA により自動正誤判定を行い、読み誤りであると推定された場合にやり直しを促す方法が考えられる。一方で報告者らは独自の CrowdSourcing 環境で作業を依頼しており、作業者に厳しい条件は必ずしも収集効率の向上には繋がらない。オンライン判定によるやり直しは作業者のモチベーション低下につながり、最悪の場合、音声の収録作業や CrowdSourcing の作業全体が嫌厭される可能性がある。

本報告では、FA によるオンラインの自動正誤判定を利用した音声の採点システムを作成し、採点システムを導入した音声収集により収集精度の向上を図る。図 3 は採点システムを導入した音声収録の作業画面を表している。作業者が録音を終了すると音声はサーバに転送され、サーバ上で FA により採点が行われ、採点結果が図 3 右のカラーバーに反映される。バーは採点結果により下から上に色が明るくなるようになっており、採点結果が閾値を超えた場合には、ボーナス報酬が加算される。採点は 3 章で導入した異常音素長率 r_{out} をもとに算出され、今回の実験では正常な音素長の割合 $r_{in} = 1.0 - r_{out}$ を用いて算出する。FA によるオンライン判定の結果を点数に変換する

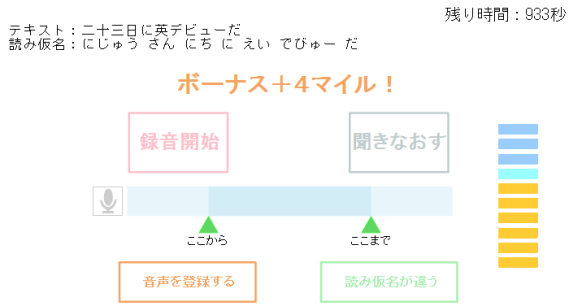

図 3: 採点あり音声収集の作業画面の例

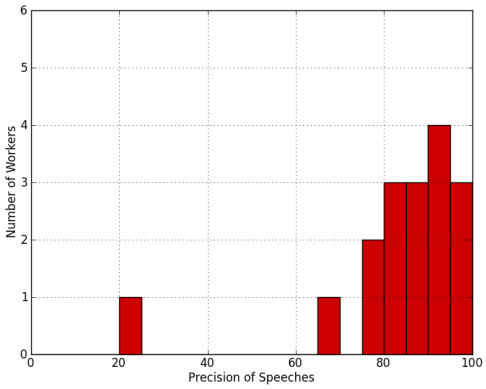

図 4: 読み上げ正解率ヒストグラム (採点あり)

ことで、低品質音声の拒絶するのではなく、高品質音声の推奨に利用することができる。これは問題に対してゲーム性のある要素を取り入れるゲーミフィケーションの一例と考えられる。

図 4 は、採点システムあり音声収集における作業者ごとの正解率のヒストグラムを表している。図のフォーマットは図 2 と同一である。採点システムあり音声収集では、全作業者の平均正解率は 87.8% となり、採点システムのなかった音声収録の平均正解率 70.9% に対し、16.9 ポイント精度が向上している*1。図 2 と図 4 を比較すると全体的に読み上げの精度が向上しており、採点によるフィードバックとボーナス報酬が、作業者の読み直しに対するストレスの低減や集中力の向上につながっていると推測される。収集時期や作業者が異なるため単純な比較はできないものの、採点システムによるゲーミフィケーションにより作業の精度が向上した一例であると考えられる。

7. まとめ

本報告では、CrowdSourcing を利用した音声収集を行い、読み誤りフィルタリングによる品質管理と、音声認識エンジンの評価音声としての実利用について報告した。また自動読み誤りフィルタリング手法を適用することで採点システムを構築し、採点システムによるゲーミフィケーションにより収集精度が向上する例を報告した。今後は、本報告の内容をもとに収録音声の学習データへの利用を検討するとともに、ゲーミフィケーションの導入を推進して作業精度の向上を図る。

*1 採点システムなしと採点システムありの音声収集には 6 か月ブランクがあり、作業者の作業への慣れの影響は小さいと考えられる。

参考文献

[Ashikawa 12] 芦川将之 and 西山修 and 下郡信宏: Crowd-Sourcing 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, 111(447)(2012), no. 4, pp. 367-381 (2011)

[Parent 11] Parent, G and Eskenazi, M: Speaking to the Crowd: looking at past achievements in using crowd-sourcing for speech and predicting future challenges, Proc. of INTERSPEECH, 3037–3040 (2011)

[Ledlie 10] Ledlie, J. and Odero, B. and Minkov, E. and Kiss, I. and Polifroni, J.: Crowd translator: on building localized speech recognizers through micro-payments, ACM SIGOPS Operating Systems Review, vol43, issue4, pp. 84-89 (2010)

[Lane 10] Lane, I. and Waibel, A. and Eck, M. and Rottmann, K.: Tools for collecting speech corpora via Mechanical-Turk, Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 184-187 (2010)

[Polifroni 10] Polifroni, J. and Kiss, I. and Seneff, S.: Speech for content creation, International Journal of Mobile Human Computer Interaction (IJMHCI), vol3, issue2, pp. 35-49 (2011)

[McGraw 10] McGraw, I. and Lee, C.Y, and Hetherington, L. and Seneff, S. and Glass, J.: Collecting voices from the cloud, Proc. LREC, vol100, (2010)

[McGraw 09] McGraw, I. and Gruenstein, A. and Sutherland, A.: A self-labeling speech corpus: Collecting spoken words with an online educational game, Proceedings Interspeech, (2009)

[Chernova 10] Chernova, S. and Orkin, J. and Breazeal, C.: Crowdsourcing hri through online multiplayer games, Proc. Dialog with Robots: AAAI fall symposium, (2010)

[Freitas 10] Freitas, J. and Calado, A. and Braga, D. and Silva, P. and Dias, M.: Crowdsourcing platform for large-scale speech data collection, Proc. FALA, Vigo, (2010)

[McGraw 11] McGraw, I. and Glass, J. and Seneff, S.: Growing a Spoken Language Interface on Amazon Mechanical Turk, Proc. Interspeech2011, (2011)

[Marge 10] Marge, M. and Banerjee, S. and Rudnicky, A.I.: Using the Amazon Mechanical Turk for transcription of spoken language, Acoustics Speech and Signal Processing (ICASSP), pp. 5270-5273 (2010)

[Li 10] 李清宰 and 河原達也: Amazon Mechanical Turk を用いた音声データ収集による音声検索システムの評価, 情報処理学会研究報告. SLP, 音声言語情報処理, vol2011, issue9 pp. 1-6 (2011)