E-012

音声ウェブシステムを用いて収集した実環境子供発話に関する調査

Investigations of Real Environmental Child Speech Collected by Voice Web System

栗原 理沙 † 西村 竜一 † 宮森 翔子 † 河原 英紀 † 入野 俊夫 †
Lisa Kurihara Ryuichi Nisimura Shoko Miyamori Hideki Kawahara Toshio Irino

1．はじめに

我々はこれまで、発話を入力とする子ども利用者の判別システムについて検討を行ってきた。その研究過程において、研究資料としての発話データベースを整備する必要があった。そこで、我々は音声ウェブシステムを用いて、家庭等の実環境における子どもの声(一部、大人の声も含む)の大規模な収集を行ってきた。これまでに収集した子ども発話の数は、3,489 個になる [1)]。しかし、収集した発話の中には、「雑音が多く十分な SN が確保されていない」、「過多な背景雑音が含まれている」等を理由に、研究資料として利用が困難なものが含まれている。

そこで、本研究では収集した発話の研究資料としての有効性を検証するため、その内容の詳細な確認・調査を行った。まず、手作業で収集発話を分類した。続いて、音声認識プログラムの強制アライメント機能を用いて、子ども発話の時間区間切り出しを行った後、SN 比を算出した。本稿では、その結果を報告する。

2．音声ウェブシステムを用いた発話収集とその問題点

今回、音声収集には音声ウェブシステム w3voice[2),3),*1] を用いた。これにより、発話者は家庭等からインターネットを通じて、録音作業に参加することができる。発話者はウェブブラウザ上で動作するインタフェースで音声の録音を行う。録音に用いた PC やマイクは、発話者自身が用意したものである。なお、発話者は楽天リサーチ社のモニタ誘引サービスを介して募集した。

今回のデータ収集では、発話者は 3 回の録音作業を完了した後、発話者の属性及び使用機材に関するアンケートに回答することが求められる。その項目の一部は以下の通りである。

・性別 ・年齢 ・録音環境 ・使用 PC とマイクのタイプ

ただし、低年齢の発話者は、上記アンケートに回答することが困難である。この場合は、保護者に録音及びアンケートの補助に協力していただくよう事前に依頼をした。

各アンケート項目の回答は、発話者(もしくは保護者)の自己申告に基づくものである。ここで問題となるのは、ウェブシステムを介している以上、その回答内容に信頼性が確保されていないことにある。そこで、すべての録音データを作業者 2 名(大学生)が耳で聞き、その中身を確認する作業を行った。以下では、上記アンケート項目が妥当に入力されていると、作業者によって判断された発話者のデータを利用可(valid)、アンケートに入力ミスが存在する、あるいは発話者が入力を偽っていると判断できるものを利用不可(invalid)と定義した。

† 和歌山大学システム工学部

*1 http://w3voice.jp/

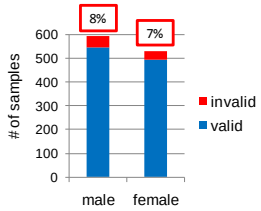

図 1 発話者の性別

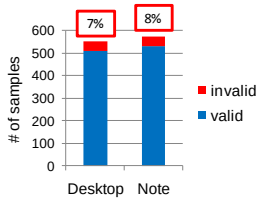

図 2 発話者が使用した PC タイプ

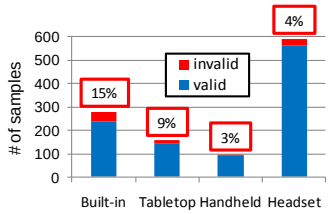

図 3 発話者が使用したマイクタイプ

2.1 発話者属性及び機材に関する調査結果

本節では各アンケートの回答結果を、前述のデータ利用可/不可の観点からまとめる。

発話者の性別分布を図 1 に示す。横軸は発話者の性別、縦軸は各性別の発話者数である。各棒グラフの青は発話が利用可能と判別された数、赤は利用不可の数を示す。図 1 からわかるように、男女ともに 90%以上のデータが利用可であり、また男女間で利用可能割合の差は無かった。

次に、録音に使用した PC タイプの結果を図 2 に示す。この項目においても、デスクトップとノート型との間に利用可否の割合に差は生じていない。

最後に、録音に使用したマイクのタイプに関する結果を図 3 に示す。アンケート回答結果から、実験に用いられたマイクはヘッドセット型のものが多く、次にパソコン内蔵タイプであることがわかった。これまでと同様、利用可/不可の割合に違いは確認できなかった。

以上の結果から、発話者属性及び使用機材と収集発話の利用可否との間には特別な関係が認められないことがわかった。つまり、今後、音声ウェブシステムを用いた発話収集を継続する場合も、発話者に特定の条件を課す必要がないと考えることができる。

3. 騒音環境に関する調査

ここまでの調査に加え、収集発話の利用可否を左右する要因に、録音時の騒音環境の影響が考えられる。ここからは、発話者が主観的に感じた騒音状態(静か/うるさい)と、SN 比に基づく客観的な騒音状態の 2 点について考察する。

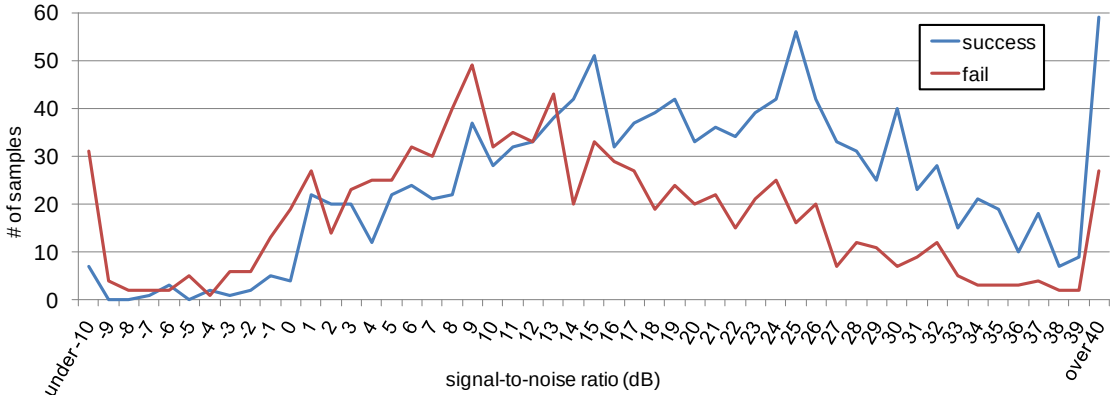

図 5　子供発話のＳＮ比分布

3.1 主観的な騒音状態に関する考察

発話者からは、アンケートの項目として、録音場所(自宅/自宅以外の屋内/屋外)及び、その場所が主観的に(静か/うるさい)の回答を得ている。各選択の発話者数と、その割合を表 1 に示す。ただし、今回、屋外で録音した発話者は、存在しなかったので省略する。

利用可能な収集発話のほとんどは、静かな自宅で録音されていた。録音場所を自宅のみに限定すると、静かな録音環境の方が、利用可能な発話が多いことが確認できる。これより、発話者が主観的に感じる環境が、データの利用可否に影響することを示唆する結果となった。

3.2　自動区間切り出しに基づく SN 比の算出

客観的な騒音状態を示す指標として、SN (Signal-to-Noise)比を用いる。はじめに、SN 比の算出手順について述べる。まず、収集した子供発話を用いて、子供発話区間の自動区間切り出しを行った。子供発話区間とその他の区間(無音区間/大人発話区間/雑音区間)を手動で定義した簡易ラベルを用意した後、音声認識器による強制アライメントを適用し、フレーム単位の時間ラベルを得た。強制アライメントには、HTK 3.4.1 に含まれる Hvite[4]を用いた。また、音響モデルには、大人男性、大人女性、子ども、非音声の 4 つのクラスを持つ GMM(Gaussian Mixture Model)を用いた。GMM の学習には、収集発話のうち 1,109 個を用いた。これら発話に関しては、音声/非音声の時間区間ラベルを作業者の手作業により作成した。GMM 学習に使用した音響特徴量は、MFCC と ΔMFCC、Δパワーである。

自動区間切り出しで分けられた音声区間と雑音区間の振幅から式(1)に基づき SN 比を求めた。

$$SNR = 20 * \log_{10} \frac{S_{RMS}}{N_{RMS}} \qquad (1)$$

3.3 客観的な騒音状態（SN 比）に関する考察

子供発話部分の自動区間切り出しの結果を目視により確認したところ、1,251 の発話が切り出しに成功し、942 の発話が失敗した。また、成功した発話(success)と失敗した発話(fail)の SN 比のヒストグラムを図 5 に示す。図の横軸は SN 比、縦軸はサンプル数である。各折れ線グラフの青は自動区間切り出しが成功した発話を、赤は自動区間切り出しが失敗した発話を示している。自動切り出しに成功した発話は、SN 比のピークが 25dB であり、平均は 19.5dB、分散は 135.2 であった。一方、自動切り出しに失敗した発話は、SN 比のピークが 9dB であり、平均は 12.8dB、分散は 237.2 であった。これより、SN 比の値が良いと、自動区間切り出しが成功することが多く、騒音状態がデータ利用可否に深く関わってくることがわかる。

表 1　発話収集実験の録音環境

録音場所	録音環境	利用可	利用不可
自宅	静か	900(93%)	64(7%)
	うるさい	126(87%)	18(13%)
自宅以外の屋内	静か	12(86%)	2(14%)
	うるさい	2(100%)	0(0%)

4．まとめ

本研究では、音声ウェブシステムを用いて収集した実環境発話の詳細な確認を行った。アンケートの分析結果より、発話者属性及び使用機材と収集発話の利用可否との間には、特別な関係が認められなかった。今後、音声ウェブシステムを用いた発話収集を行う場合も、発話者に厳密な実験条件を課す必要がないと考えることができる。また、主観的な騒音状態に関して、発話者が静かだと感じる環境であれば、利用可能な発話データを得られることがわかった。さらに、SN 比の値が良い発話データは、有用性が高いことも確認した。

今後は、今回得られた知見に基づき音声ウェブシステムを用いた発話収集実験のデザインを再検討し、より効率の良い発話収集方法を提案したいと考えている。

謝辞　本研究の一部は、科学研究費補助金及び和歌山大学平成 22 年度学長裁量経費の支援を受けた。

5．参考文献

1) 宮森 他, "ウェブ収集発話を対象とした若年者判別の検討", 情報処理学会創立 50 周年記念（第 72 回)全国大会講演論文集, vol.2 pp.285-286, 5U-7, 2010.
2) 西村 他, “音声入力・認識機能を有する Web システム w3voice の開発と運用”, 情報処理学会研究報告, 2007-SLP-68-3, 2007.
3) Ryuichi Nisimura, et al., "Development of Speech Input Method for Interactive VoiceWeb Systems", Lecture Notes in Computer Science, vol.5611, pp.710-719, Springer, 2009.
4) Steve Young, et al., ”The HTK Book (for HTK Version 3.4)”, Cambridge University Engineering Department., 2006.