声の主観的評価と音声特徴量の相関分析

Correlation Analysis Between Subjective Voice Evaluation and Acoustic Features

寺下逸生テオ [1]　日向寺拓海 [1]　川勝真喜 [1]
ItsukiTheo Terashita　Takumi Hyugaji　Masaki Kawakatsu

東京電機大学　システムデザイン工学部　情報システム工学科 [1]
Department of Information System Engineering, School of System Design and Technology, Tokyo Denki University

1. はじめに

現在,音声読み上げ技術やスマートアシスタントの普及により,機械音声を耳にする機会が増加している.しかし,一部の機械音声は聞き取りにくく,ユーザー体験を損なう要因となっている.

本研究では，音声の聞き取りやすさや好みに影響する要因を明らかにするため，8 人の話者による同一文章の音声を 100 人の被験者に聴取させて評価を行った．被験者による主観的な評価結果と使用した音声の解析結果の相関分析を通じて,聞き取りやすさに寄与する要因を考察・検討した．これにより,聞き取りやすい機械音声の設計指針を提供し,音声技術の向上に貢献することを目指す.

2. 実験

高道らによる公開データセット[1]を用いて被検者 100 人(男性：90 人,女性：10 人,年齢 18~54 才,平均年齢:21.32 才)を対象に防音室で主観評価実験を行った.スピーカーは被検者から前方 1.6 m,左右で 1 m の間隔をあけて配置した．被検者はスピーカーが頭の位置の正面に来るように長椅子に着席させ,1 度の実験につき 1~4 人で行った.

また，8 つの音声に加えて，3 つのサンプル音声を含む計 11 種類の音声を使用した．サンプル音声は，提示される文章の内容を聴かせる目的で使用されるものであり，解析には含めていない．音声はランダムな順番で聴取させた．実験構成は,説明・質疑：約 5 分,サンプル音声再生：約 1 分,音声再生・設問回答：約 16 分,実験後感想回答：約 2~3 分以上の計約 25 分で構成された． (図 1)

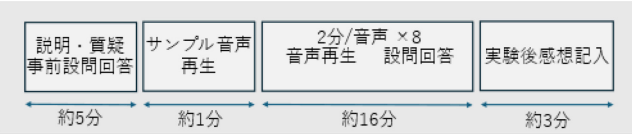

図 1　実験手順

設問の内容は,文献[2]の調査項目を基に,声色に関わる問いとしゃべり方に関する問いに加え,「この声は好みか」という 1~5 の 5 段階評価,さらに「年齢」,「性別(生物学的)」,「直近 2 年以上の音楽歴」の 18 問で構成した．得られた結果から平均値を算出した.

各話者の音声データから特徴量を算出した．具体的には,音声データを 0.1 秒ごとにウィンドウ化し,各ウィンドウにおいて基本周波数,音の大きさ,MFCC（1～13 次元）の平均値および標準偏差を抽出した.なお,ウィンドウには無音部分が含まれないよう設定した．また,無音部分を話者の間と見なし,その合計時間および標準偏差を算出した．さらに,有音期間の総時間も特徴量として算出した.

3. 結果

各設問の平均値と特徴量との相関係数を算出し,上位 3 項目を表 1 に示した．表 1 に示される変数は,以下のように定義される．silent (sl)_time は無音秒数,active_time は発話秒数,db は音の大きさ,mfcc_***i***は MFCC の第***i***次元,freq は基本周波数（f0）を表し,これらは変数の前半部分を構成する．また,mean および std はそれぞれフレームごとの平均値と標準偏差を示し,変数の後半部分を構成する.

表 1　集計結果と特徴量の相関係数の上位 3 項目

設問項目 ＼ 順位	1st	2nd	3rd
かすれ度合	sl_time_std (0.771)	silent_time (0.737)	mfcc_8_std (0.710)
この声は好みか	mfcc_3_mean (0.767)	mfcc_10_mean (0.693)	mfcc_4_mean (-0.64)
ボリューム	mfcc_13_mean (0.856)	db_maen (0.741)	mfcc_1_mean (0.729)
リッラクスの度合い	mfcc_3_mean (0.754)	mfcc_10_mean (0.714)	mfcc_6_std (-0.693)
温かさ	mfcc_13_std (0.814)	mfcc_3_mean (0.684)	mfcc_3_std (-0.632)
響き	mfcc_10_mean (0.874)	mfcc_3_mean (0.855)	mfcc_12_mean (0.529)
好みのイントネーションか	mfcc_3_mean (0.690)	mfcc_4_mean (-0.654)	mfcc_10_mean (0.650)
好みのことばの間	mfcc_1_std (-0.917)	mfcc_3_std (-0.679)	mfcc_7_std (0.650)
高さ	freq_mean (0.974)	mfcc_5_mean (-0.963)	mfcc_8_mean (-0.892)
柔らかさ	mfcc_11_mean (-0.839)	mfcc_12_std (-0.819)	mfcc_13_mean (-0.719)
心地よさ	mfcc_10_mean (0.825)	mfcc_11_mean (-0.781)	mfcc_3_mean (0.735)
速度	sl_time_std (-0.986)	active_time (-0.933)	mfcc_11_std (0.791)
明るさ	mfcc_2_mean (-0.831)	freq_std (0.816)	jitter_std (0.797)
明瞭であるか（滑舌）	mfcc_8_std (-0.846)	mfcc_2_mean (-0.817)	mfcc_11_mean (-0.640)
力強さ	mfcc_13_mean (0.743)	mfcc_11_std (0.640)	mfcc_9_std (0.637)

結果より,ボリューム,温かさ,響き,好みのことばの間,柔らかさ,心地よさ,明るさ,明瞭であるか(滑舌)の項目について高い相関を示す特徴量が存在することが明らかとなった．また今回,5 段階の設問項目は被検者ごとに尺度の差が生まれていた可能性があるため,今後被検者ごとに正規化した集計結果を使用する方法を検討する.

参考文献

[1] S. Takamichi, et. al. ″JVS corpus: free Japanese multi-speaker voice corpus,″ arXiv preprint, 1908.06248, Aug. 2019.

[2] W.Benjamin, et. al. ″Voice attributes affecting. Likability perception.″, Interspeech, 2010.