Age Classification of Icelandic Speech using Sequential Feature Elimination

Abstract

We propose an age classifier for Icelandic speech utilizing the Samrómur dataset. Through sequential feature elimination, we select 25 features out of 6,378 Opensmile features, achieving an F1 score of 0.824. This method progressively removes less informative features, prioritizing the most discriminative ones. Our findings indicate that these 25 features capture physiological changes associated with aging (e.g., lower pitch) and generational speech habits (e.g., energy variations), underscoring their relevance to age-related speech patterns.

1 Introduction

The human voice undergoes noticeable transformations as individuals progress through different stages of life. These changes manifest in various aspects of speech production (Mueller, 1997; Schultz et al., 2021), and are due to changes in anatomic structure, physiology, and motor control (Stathopoulos et al., 2011). Children have shorter vocal tracts and smaller vocal folds than adults, resulting in a higher fundamental frequency (F0) value and in higher-pitched tonalities (Gautam and Singh, 2018; Banik et al., 2015; Remacle et al., 2020; Lopes et al., 2014). In contrast, individuals aged 60 and above often demonstrate slower speech rate and decreased vocal energy. Women develop a lower-pitched voice, while men develop a higher-pitched voice. Common to both sexes is a breathier, shakier and more strained voice (Smiljanic and Gilbert, 2017; Stathopoulos et al., 2011; Tarafder et al., 2012; Kaneko et al., 2015; Martins et al., 2015).

Accurately classifying voices into different age categories holds practical significance in various domains, ranging from speech recognition systems to personalized healthcare.

Automatic speech recognition (ASR) systems rely on acoustic features to accurately transcribe spoken language. By incorporating age-specific models or adapting the recognition process based on the detected age group, these systems could likely achieve a significantly reduced error rate.

Furthermore, voice-controlled personal assistants and virtual agents can leverage age classification to provide tailored responses and personalized experiences. Recognizing the age group of the user allows these systems to adjust their behavior, tone, and language to match the preferences and needs of different age demographics. For instance, an age classifier can enable a voice assistant to offer age-appropriate content recommendations, educational materials, or entertainment options.

In addition, age classification in speech has potential applications in healthcare. Monitoring and analyzing the speech patterns of elderly individuals can aid in the early detection of age-related vocal pathologies, such as dysphonia or vocal cord disorders. By establishing baseline vocal characteristics specific to different age groups, clinicians and researchers can better track changes in voice quality over time and identify potential health issues.

In this paper, we present the result of a voice classification experiment to classify voices as belonging to a child, adult or elderly individual. This study is based on Icelandic, a North Germanic language spoken by approximately 370'000 people.

2 Background

Recent studies have demonstrated the potency of leveraging Convolutional Neural Networks (CNN) in combination with Temporal Convolutional Networks (TCN) for reliable age group classification in Interactive Voice Response (IVR) systems, even amidst data balance challenges (Sánchez-Hevia et al., 2022). Concurrently, advancements are being made in employing a variety of neural network architectures, transfer learning approaches, and multitask learning techniques. Notably, a two-stage transfer learning scheme with a QuartzNet embed-

der has exhibited breakthrough performance in age estimation tasks, while a d-vector-based system has proven robust for gender classification (Kwasny and Hemmerling, 2021). Additionally, the emergence of a novel CNN model integrated with a Multi-scale Attention Mechanism (MAM) for handling age, gender, and age-gender classification tasks represents a promising development. Despite its high accuracy scores, the MAM mechanism grapples with differentiating closely related age groups, such as teens and twenties, highlighting the need for further research (Tursunov et al., 2021). These studies collectively illuminate the nuanced complexities and vast potential of deep-learning applications in speech-based age and gender recognition, heralding the advent of more sophisticated and accurate speech-processing systems.

To achieve their impressive performance, such models require intricate architectures. However, their complexity is traded off for interpretability. In fact, deep-learning models are notorious for their difficult-to-interpret results. Our goal in this paper is to uncover the most informative features in the dataset rather than pushing the performance of the state-of-the-art. In our study, we have included a Multilayered Perceptron (MLP), which is the simplest form of deep-learning model, for comparison. For feature selection, we have used a Random Forest classifier, arguably one of the most interpretable yet powerful classifier architectures.

3 Methods

We have built an age group classifier on the Samrómur dataset [1], which contains recordings in Icelandic of 27795 individuals aged 4 to over 90. The speakers read pre-written text out loud. We have featurized the dataset using openSMILE (Eyben, 2015), a software for audio-feature extraction offering 6,378 features focusing on various properties of sound waves. Some features analyze the frequency by simple statistics such as mean, standard deviation, minimum, maximum, percentiles, etc. Other features measure spectral (i.e., in the frequency domain) information, such as the rate of change and distribution of frequency across the spectrum. Many of the features come from psychoacoustic filterbanks, i.e. sound filters designed to simulate the manner in which the human ear and brain process sound. As such, openSMILE provides us with a comprehensive and detailed battery of tools to analyze speech samples when determining the age of speakers.

The speaker ages were divided into 3 bins - 4–19, 20–59, and 60+, and the features of each audio segment were extracted as a single vector. There were 4,613, 11,590 and 1,099 recordings in each of the bins, respectively, for a total of 17,302.

To predict the age group from the features, we have applied a selection of algorithms and hyperparameter configurations, as well as sequential feature selection, a feature-elimination method that determines the smallest group of features needed to make a prediction without a statistically significant reduction in performance. The method iteratively evaluates and eliminates features based on their importance or performance. It starts with an initial set of features and progressively selects or removes features until we obtain the minimal subset of features whose removal significantly reduces performance. The goal is to identify a subset of features that are most relevant to the task at hand while potentially reducing complexity.

To compare the different classification methods, we have used the weighted F1 score, which combines recall (number of true positives) with accuracy (calculated as (true positives + true negatives) over total instances) into a single score that takes both of them into account. It is meant to address the trade-off between precision and recall in a single metric, providing a balanced evaluation of a classification model's performance.

For each model class, we have compared a selection of hyperparameter combinations using k-fold cross-validation on the train set. The top-performing combination was evaluated on the evaluation set.

4 Results

Table 1 shows the F1 score, alongside precision and recall, for each of the top hyperparameter combinations of each model class. As can be seen, the MLP classifier achieved the best performance.

After the feature elimination process, the Random Forest classifier identified 25 final features whose elimination leads to a significant reduction in performance (meaning the same is untrue of all other features). They are given in Table 2, ordered by importance, alongside a description of what they measure.

[1] Samrómur is an output of the Icelandic Language Technology Programme (Nikulásdóttir et al., 2020, 2022); See: https://samromur.is/gagnasafn (Mollberg et al., 2020; Mena et al., 2022; Hedström et al., 2022)

Classifier	F1 Score	Precision	Recall
MLP	**.824**	**.80**	**.81**
Random Forest	.74	.76	.78
K-Means	.71	.75	.76

Table 1: Classifier Performance (weighted by class)

5 Discussion

The identified features seem to primarily focus on the spectral content and prosody of the speech signal, rather than on aspects such as rhythm, pitch, or timbre. The selection heavily emphasizes objective spectral and energy characteristics, while psychoacoustic features that approximate perceived frequency and quality (and hence associated with pitch and timbre) are relatively less represented. In detail:

1. `F0final_sma_quartile1` and `F0final_sma_percentile1.0`: These are measures of pitch distribution in the speech signal, representing the first quartile and the first percentile of the fundamental frequency (F0) distribution respectively. They could reflect the lowest pitch values in the speaker's voice.

2. MFCC-related features: The Mel Frequency Cepstral Coefficients (MFCCs) capture the spectral shape of the sound and are used extensively in speech and audio processing. They could provide information about the speaker's vocal tract shape and size.

3. `pcm_fftMag_spectralRollOff25.0_sma_percentile1.0` and `pcm_fftMag_spectralRollOff50.0_sma_percentile1.0`: These are measures of the distribution of spectral energy. Spectral roll-off points are the frequencies below which a certain percentage of the total spectral energy lies. These could indicate the concentration of energy in lower-frequency components.

4. `audSpec_Rfilt_sma*_de[4]_peakRangeAbs`, `audSpec_Rfilt_sma[4]_lpc2`, `audSpec_Rfilt_sma[4]_range`, `audSpec_Rfilt_sma[4]_lpc1`, etc.: These features, related to the auditory spectrum and linear predictive coding (LPC), could provide information about the speech signal's spectral envelope, which is affected by the speaker's vocal tract characteristics.

5. `pcm_RMSenergy_sma_lpc1`: This is a measure of the speech signal's energy, which can be related to the speaker's loudness or vocal effort.

6. `pcm_fftMag_spectralKurtosis_sma_de_meanSegLen` and `audSpec_Rfilt_sma[0]_maxSegLen`: These could capture aspects of the distribution of spectral energy across different frequency bands and speech segments.

It is possible that some of these features highlight physiological changes while others highlight psychological differences. As explained above, some of the features directly measure properties of the soundwave that are due to vocal-fold vibrational frequency and physiological characteristics of the vocal tracts (for example MFCC-related and $F0_{final}$ features). On the other hand, features that measure frequency distribution across the soundwave (that is, what percentage of the total spectral energy lies between which cutoff points, as opposed to what is the absolute average frequency), such as magnitude spectral roll of features, are more likely to correspond to differences in psychological attitude between speakers.

6 Conclusions

In this study, we developed an age classifier for Icelandic speech using the Samrómur dataset. Through sequential feature elimination, we identified 25 informative features out of 6,378 Opensmile features, achieving an F1 score of 0.824. These features encompassed various aspects of the speech signal, including frequency distribution, vocal tract characteristics, energy levels, and spectral envelope. They potentially capture physiological changes associated with aging and psychological differences in speech patterns. To explore this possibility in more detail, future studies should incorporate an in-depth statistical analysis of the interaction of each feature class with age in order to break down its effect size.

The accurate classification of age in speech has practical implications in speech recognition systems, voice-controlled assistants, and healthcare applications. Our findings shed light on the significance of specific features in capturing age-related speech patterns.

Future research should focus on investigating the generalizability of the selected features across

Feature	Description
F0final_sma_quartile1	Quartile 1 of F0 final
mfcc_sma[13]_rqmean	RQ mean of MFCC [13]
F0final_sma_percentile1.0	Percentile 1.0 of F0 final
pcm_fftMag_spectralRollOff25.0_sma_percentile1.0	Percentile 1.0 of PCM FFT MSR 25.0
mfcc_sma[4]_quartile1	Quartile 1 of MFCC [4]
mfcc_sma[13]_percentile1.0	Percentile 1.0 of MFCC [13]
mfcc_sma[13]_stddev	Standard deviation of MFCC [13]
audSpec_Rfilt_sma_de[4]_peakRangeAbs	Peak range absolute of auditory spectrum Rfilt de[4]
pcm_fftMag_spectralRollOff50.0_sma_percentile1.0	Percentile 1.0 of PCM FFT MSR 50.0
mfcc_sma[13]_quartile2	Quartile 2 of MFCC [13]
audSpec_Rfilt_sma[4]_lpc2	LPC2 of auditory spectrum Rfilt [4]
audSpec_Rfilt_sma[4]_range	Range of auditory spectrum Rfilt [4]
audSpec_Rfilt_sma[4]_lpc1	LPC1 of auditory spectrum Rfilt [4]
audSpec_Rfilt_sma[1]_lpc1	LPC1 of auditory spectrum Rfilt [1]
pcm_RMSenergy_sma_lpc1	LPC1 of PCM RMS energy
audspec_lengthL1norm_sma_de_lpc2	L1-norm length of auditory spectrum de LPC2
mfcc_sma[4]_stddev	Standard deviation of MFCC [4]
audspecRasta_lengthL1norm_sma_lpc2	L1-norm of RASTA-filtered aud. spec. LPC2
mfcc_sma_de[1]_meanRisingSlope	Mean rising slope of MFCC de[1]
mfcc_sma_de[1]_range	Range of MFCC de[1]
audSpec_Rfilt_sma_de[4]_meanFallingSlope	Mean falling slope of auditory spectrum Rfilt de[4]
audSpec_Rfilt_sma[0]_maxSegLen	Maximum segment length of aud. spec. Rfilt [0]
pcm_fftMag_spectralKurtosis_sma_de_meanSegLen	Mean seg. length of PCM FFT mag. spec. kurtosis
mfcc_sma[1]_peakMeanAbs	Peak mean absolute of MFCC [1]
mfcc_sma[1]_range	Range of MFCC [1]

Table 2: Top performing audio features. SMA = Spectral Shape Mean Amplitude. PCM = Pulse Code Modulation, a standard method for representing analog signal digitally. LPC = Linear Predictive Coding, a linear time-series model of speech signals. MFCC = Mel-Frequency Cepstral Coefficients, a technique for approximating human auditory perception by transforming the frequency into the Mel scale, which is the perceptually relevant one. FFT = Fast Fourier Transform. RASTA (=RelAtive SpecTrAl) filter is a technique used in speech signal processing to enhance the robustness of features extracted from speech signals. MSR = Magnitude Spectral Roll-off, which provides information about the frequency below which a certain percentage of the signal's total energy is concentrated. Kurtosis is a measure of how heavy the tails of a distribution is relative to a normal distribution.

different languages and exploring their potential in detecting age-related vocal pathologies. Additionally, more advanced analytical pipelines incorporating powerful deep-learning architectures and feature extraction methods, as discussed in Section 2, should be explored to further enhance the classification accuracy and feature selection process.

7 Limitations

Our study has a number of limitations. Firstly, we employed relatively simple classification architectures, prioritizing interpretability over performance. Exploring more advanced models may yield higher classification accuracy.

Secondly, our analysis was limited to the openSMILE feature extraction pipeline and the Samrómur dataset. To enhance the generalizability and accuracy of our findings, it would be beneficial to incorporate alternative feature extraction methods, diverse datasets, and different classification architectures. Expanding the scope of our study could provide a more comprehensive understanding of age classification in speech and improve the applicability of our results across different contexts and languages.

References

Anindita Banik, Shantanu Arya, and Anjali Kant. 2015. Vocal parameters in children between 4 to 12 years of age: An attempt to establish a prototype database. *Intern. J. Scient. Research Publications*, 11:446–453.

Florian Eyben. 2015. *Real-time speech and music classification by large audio feature space extraction*. Springer.

Sumanlata Gautam and Latika Singh. 2018. Variations in relationship between fundamental frequency and intensity in speech of children during development. *Journal of Cultural Cognitive Science*, 2:59–69.

Staffan Hedström, David Erik Mollberg, Ragnheiður Þórhallsdóttir, and Jón Guðnason. 2022. Samrómur: Crowd-sourcing large amounts of data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2311–2316.

Mami Kaneko, Shigeru Hirano, Ichiro Tateya, Yo Kishimoto, Nao Hiwatashi, Masako Fujiu-Kurachi, and Juichi Ito. 2015. Multidimensional analysis on the effect of vocal function exercises on aged vocal fold atrophy. *Journal of Voice*, 29(5):638–644.

Damian Kwasny and Daria Hemmerling. 2021. Gender and age estimation methods based on speech using deep neural networks. *Sensors*, 21(14):4785.

Leonardo Wanderley Lopes, Ivonaldo Leidson Barbosa Lima, Elma Heitmann Mares Azevedo, Maria Fabiana Bonfim de Lima-Silva, Débora Pontes Cavalcante, Larissa Nadjara Alves de Almeida, and Anna Alice Figueirêdo de Almeida. 2014. Vocal characteristics during child development: perceptual-auditory and acoustic data. *Folia Phoniatrica et logopaedica*, 65(3):143–147.

Regina Helena Garcia Martins, Adriana Bueno Benito Pessin, Douglas Jorge Nassib, Anete Branco, Sergio Augusto Rodrigues, and Selma Maria Michelim Matheus. 2015. Aging voice and the laryngeal muscle atrophy. *The Laryngoscope*, 125(11):2518–2521.

Carlos Daniel Hernandez Mena, David Erik Mollberg, Michal Borskỳ, and Jón Guðnason. 2022. Samrómur children: An icelandic speech corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 995–1002.

David Erik Mollberg, Ólafur Helgi Jónsson, Sunneva Þorsteinsdóttir, Steinþór Steingrímsson, Eydís Huld Magnúsdóttir, and Jón Guðnason. 2020. Samrómur: Crowd-sourcing data collection for icelandic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3463–3467.

Peter B Mueller. 1997. The aging voice. In *Seminars in speech and language*, volume 18, pages 159–169. © 1997 by Thieme Medical Publishers, Inc.

Anna Björk Nikulásdóttir, Þórunn Arnardóttir, Starkaður Barkarson, Jón Guðnason, Þorsteinn Daði Gunnarsson, Anton Karl Ingason, Haukur Páll Jónsson, Hrafn Loftsson, Hulda Óladóttir, Eiríkur Rögnvaldsson, et al. 2022. Help yourself from the buffet: National language technology infrastructure initiative on clarin-is. In *CLARIN Annual Conference*, pages 109–125.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for icelandic 2019-2023. *arXiv preprint arXiv:2003.09244*.

Angélique Remacle, Ysaline Genel, Magali Segers, and Marc De Bodt. 2020. Vocal characteristics of 5-year-old children: proposed normative values based on a french-speaking population. *Logopedics Phoniatrics Vocology*, 45(1):30–38.

Héctor A Sánchez-Hevia, Roberto Gil-Pita, Manuel Utrilla-Manso, and Manuel Rosa-Zurera. 2022. Age group classification and gender recognition from speech with temporal convolutional neural networks. *Multimedia Tools and Applications*, 81(3):3535–3552.

Benjamin G Schultz, Sandra Rojas, Miya St John, Elaina Kefalianos, and Adam P Vogel. 2021. A cross-sectional study of perceptual and acoustic voice characteristics in healthy aging. *Journal of Voice*.

Rajka Smiljanic and Rachael C Gilbert. 2017. Acoustics of clear and noise-adapted speech in children, young, and older adults. *Journal of Speech, Language, and Hearing Research*, 60(11):3081–3096.

Elaine T Stathopoulos, Jessica E Huber, and Joan E Sussman. 2011. Changes in acoustic characteristics of the voice across the life span: Measures from individuals 4–93 years of age.

Kamrul Hassan Tarafder, Pran Gopal Datta, and Ahmed Tariq. 2012. The aging voice. *Bangabandhu Sheikh Mujib Medical University Journal*, 5(1):83–86.

Anvarjon Tursunov, Mustaqeem, Joon Yeon Choeh, and Soonil Kwon. 2021. Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*, 21(17):5892.