Improving Speech Recognition through Automatic Selection of Age Group – Specific Acoustic Models

Annika Hämäläinen[1], Hugo Meinedo[2], Michael Tjalve[4], Thomas Pellegrini[5], Isabel Trancoso[2,3], and Miguel Sales Dias[1]

[1] Microsoft Language Development Center & ISCTE - University Institute of Lisbon, Lisbon, Portugal
[2] INESC-ID Lisboa, Lisbon, Portugal
[3] Instituto Superio Técnico, Lisbon, Portugal
[4] Microsoft & University of Washington, Seattle, WA, USA
[5] IRIT - Université Toulouse III - Paul Sabatier, Toulouse, France
{t-anhama,michael.tjalve,miguel.dias}@microsoft.com,
{hugo.meinedo,isabel.trancoso}@inesc-id.pt, pellegri@irit.fr

Abstract. The acoustic models used by automatic speech recognisers are usually trained with speech collected from young to middle-aged adults. As the characteristics of speech change with age, such acoustic models tend to perform poorly on children's and elderly people's speech. In this study, we investigate whether the automatic age group classification of speakers, together with age group –specific acoustic models, could improve automatic speech recognition performance. We train an age group classifier with an accuracy of about 95% and show that using the results of the classifier to select age group –specific acoustic models for children and the elderly leads to considerable gains in automatic speech recognition performance, as compared with using acoustic models trained with young to middle-aged adults' speech for recognising their speech, as well.

Keywords: Age group classification, acoustic modelling, automatic speech recognition, children, elderly, paralinguistic information.

1 Introduction

Currently available speech recognisers do not usually work well with children's or elderly people's speech. This is because several parameters of the speech signal (e.g. fundamental frequency, speech rate) change with age [1-4] and because the acoustic models (AMs) used for automatic speech recognition (ASR) have typically been trained with speech collected from young to middle-aged adults only, to serve mainstream business and research requirements. Furthermore, both children and elderly people are more likely to interact with computers using everyday language and their own commands, even when a specific syntax is required [5-9]. As compared with young to middle-aged adults, significantly higher word error rates (WERs) have been reported both for children [10-12] and for elderly speakers [10, 13, 14]. Improvements

in ASR performance have been reported when using AMs adapted to children [10-12] and to the elderly [10, 13, 14], respectively, and more and more children's (e.g. [15-17]) and elderly speech corpora suitable for training AMs (e.g. [17-19]) are gradually becoming available. However, in many speech-enabled applications, the age of the user is unknown in advance. So, if multiple sets of AMs tailored to different age groups are available, the optimal set must either be selected manually by the user, or an automatic method must be devised for selecting it.

A real-life example of a speech-enabled application that is used by people of widely varying ages is the Windows Phone app World Search, which allows users to perform web searches via the Bing search engine using their voice. The European Portuguese version of the app (currently the only version available), uses three sets of AMs optimised for three age groups: children, young to middle-aged adults and elderly people. The models optimised for young to middle-aged adults are used by default. However, through a setting in the application, users have the option of manually selecting the set of AMs that they think is the most appropriate for them. Using the default models in the case of children and the elderly is expected to deteriorate the ASR performance dramatically (cf. [10-14]). However, having to make a manual selection is rather cumbersome from the usability point of view. An accurate age group classifier would, on the other hand, allow the optimal set of AMs to be selected automatically. Similarly, it could be used to automatically select a language model and a lexicon that represents the typical human-computer interaction (HCI) of users belonging to a given age group. In spoken dialogue systems, an age group classifier might be useful for selecting dialogue strategies or different ways of interacting with the user. For example, the persona and verbosity of the responses could be adapted to better match the typical preferences of the age group of the active user. A more fun and engaging way of addressing the user could be used if (s)he were recognised as a child, whereas a more polite way, which the elderly might prefer in HCI (cf. [8]), could be applied in the case of the elderly.

The goal of this paper is to investigate whether the automatic age group classification of speakers, together with age group –specific AMs, could improve ASR performance. Although much research has been done on automatic age estimation (e.g. [20-22]), we are not aware of other studies that would have used the results of automatic age estimation to select age group –specific AMs for improving ASR performance. After describing the speech material in Section 2, we present our age group classifier and the results of our age group classification experiments in Section 3. Section 4 presents the automatic speech recogniser and the age group –specific AMs used in this study, together with ASR results obtained using the default models, and age group –specific models selected using the speakers' real age and automatically detected age. We present our conclusions in Section 5.

2 Speech Material

The speech material used in this study originates from four different corpora of European Portuguese: the CNG Corpus of European Portuguese Children's Speech [16] (hereafter "CNG"); the EASR Corpus of European Portuguese Elderly Speech [18] (hereafter "EASR"); the BD-PUBLICO corpus [23], which contains young to

middle-aged adults' speech; and a small corpus of European Portuguese young to middle-aged adults' speech collected in Lisbon in the summer of 2013 (hereafter "YMA").

The speakers in the CNG Corpus are 3-10 years of age, whereas the speakers in the EASR Corpus are aged 60 or over. Both corpora contain prompted speech: read or repeated speech in the case of the CNG Corpus, and read speech in the case of the EASR Corpus. They come with manually verified transcriptions, as well as annotations for filled pauses, noises and "damaged" words (e.g. mispronunciations, false starts). These two corpora were used for training and testing the age group classifier, for training AMs optimised for children's and elderly speech, as well as for the ASR experiments. While the training data extracted from both corpora contains phonetically rich sentences, different types of number expressions etc., we only used a subset of utterance types for testing purposes. Our development test set, which was used in the automatic age group classification experiments, only contained the longest utterance types because very short utterances are difficult to accurately estimate age (and other speaker characteristics) from. In the case of the CNG Corpus, the development test set included sequences of cardinal numbers and phonetically rich sentences. In the case of the EASR Corpus, they comprised phonetically rich sentences only. We only used phonetically rich sentences in the evaluation test set, both in the age group classification and in the ASR experiments. This is because we wanted to maximise the comparability of test data extracted from different corpora.

The BD-PUBLICO Corpus contains newspaper sentences read out by 18-48-year-old speakers, i.e., young to middle-aged adults. The transcriptions in this corpus have not been verified manually or annotated for noises etc. We used data from this corpus in the age classification experiments, both in the training and development test sets. We tested the age group classifier and carried out the ASR experiments using the YMA Corpus, which contains phonetically rich sentences read out by speakers aged 25-59. Each speaker in the YMA Corpus uttered 80 phonetically rich sentences, 20 of which originate from the same pool of phonetically rich sentences that were used for recording the CNG Corpus and 60 of which originate from the same pool of phonetically rich sentences that appear in the EASR Corpus. This makes the evaluation test sets used in the ASR experiments comparable across all three age groups. The transcriptions in the YMA Corpus are not verified manually nor annotated. However, the recordings were monitored closely and speakers were asked to reread sentences that they did not read correctly or that included filled pauses or noises. In the age group classification experiments, we used an additional set of speakers (hereafter "YMA-a") recorded during the YMA data collection. These speakers were left out from the final YMA Corpus because they did not record the full set of 80 utterances. However, they were useful for increasing the number of speakers aged up to 54 in the training and development test sets.

The training sets, development test sets and evaluation test sets are summarised in Tables 1, 2 and 3, respectively. Before the age group classification experiments, the data were automatically pre-processed to boost the energy levels and to remove unwanted silences. The feature extraction of the training data from the CNG and EASR corpora was carried out at a frame rate of 10 ms using a 25-ms Hamming window and a pre-emphasis factor of 0.98. 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first, second and third order time derivatives were

calculated, and the total number of features was reduced to 36 using Heteroscedastic Linear Discriminant Analysis (HLDA). These features were used for building AMs optimised for children's and elderly speech (see Section 4.1).

Table 1. The main statistics of the data used to train the age group classifier for children (CNG), young to middle-aged adults (BD-PUBLICO and YMA-a) and the elderly (EASR), and to optimise acoustic models for children (CNG) and the elderly (EASR)

	CNG	BD-PUBLICO	YMA-a	EASR
#Speakers	432	109	13	778
#Male+#Female	190 + 242	55 + 54	6 + 7	203 + 575
#Word types	605	16,517	1663	4905
#Word tokens	102,537	195,169	5688	482,208
#Utterances	18,569	9320	795	44,033

Table 2. The main statistics of the development test sets used in the age group classification experiments

	CNG	BD-PUBLICO	YMA-a	EASR
#Speakers	26	10	2	48
#Male+#Female	12 + 14	5 + 5	1 + 1	16 + 32
#Word types	480	2783	644	3492
#Word tokens	6221	16,758	1550	31,565
#Utterances	866	584	160	2836

Table 3. The main statistics of the evaluation test sets used in the age group classification and in the ASR experiments

	CNG	YMA	EASR
#Speakers	51	68	96
#Male+#Female	22 + 29	36 + 32	29 + 67
#Word types	747	4485	5728
#Word tokens	3439	46,987	49,580
#Utterances	1735	5440	5351

3 Age Group Classifier

One of the goals of our study was to develop an age group classifier for automatically determining the age group of speakers belonging to one of the following three age groups: children, young to middle-aged adults, or elderly people. To achieve this goal, we developed an age group classification approach that uses two modules. First, it extracts relevant acoustic features from the speech signal, effectively transforming and reducing the dimensionality space of the input data. Second, it tries to determine which output class (i.e. age group) the speech input belongs to. The following subsections present the feature extraction frontends and the age group classification experiments, and discuss the results obtained.

Table 4. The acoustic feature set used in the age group classifiers: 65 Low Level Descriptors (LLDs)

4 energy related LLD	Group
Sum of auditory spectrum (loudness)	prosodic
Sum of RASTA-filtered auditory spectrum	prosodic
RMS Energy, Zero-Crossing Rate	prosodic
55 spectral LLDs	Group
MFCC 1-14	cepstral
RASTA-filtered auditory spectrum	spectral
Spectral energy 250–650 Hz, 1 k–8 kHz	spectral
Spectral Roll-Off Pt. 0.25, 0.5, 0.75, 0.9	spectral
Spectral Flux, Centroid, Entropy, Slope	spectral
Psychoacoustic Sharpness, Harmonicity	spectral
Spectral Variance, Skewness, Kurtosis	spectral
6 voicing related LLDs	Group
F0 (SHS & Viterbi smoothing)	prosodic
Voicing Probability	voice quality
log HNR, Jitter (local & δ), Shimmer (local)	voice quality

Table 5. The acoustic feature set used in the age group classifiers: Statistic functionals applied to the LLDs

Functionals applied to LLD / Δ LLD	Group
quartiles 1–3, 3 inter-quartile ranges	percentiles
1 % percentile (≈ min), 99 % pctl. (≈ max)	percentiles
percentile range 1 %–99 %	percentiles
position of min / max, range (max – min)	temporal
arithmetic mean, root quadratic mean	moments
contour centroid, flatness	temporal
standard deviation, skewness, kurtosis	spectral
rel. dur. LLD is above 25/50/75/90% range	temporal
relative duration LLD is rising	temporal
rel. duration LLD has positive curvature	temporal
gain of linear prediction (LP), LP Coeff. 1–5	modulation
Functionals applied to LLD only	Group
mean value of peaks	peaks
mean value of peaks – arithmetic mean	peaks
mean / std.dev. of inter peak distances	peaks
amplitude mean of peaks, of minima	peaks
amplitude range of peaks	peaks
mean / std. dev. of rising / falling slopes	peaks
linear regression slope, offset, quadratic error	regression
quadratic regression a, b, offset, quadratic error	regression

3.1 Feature Extraction

We extracted features from the speech signal using TUM's open-source openSMILE toolkit [24]. This feature extraction toolkit is capable of producing a wide range of acoustic speech features and has been used in many paralinguistic information and speaker trait detection tasks [25, 26]. The feature set used in our age group classifier

contains 6015 static features obtained by applying statistic functionals to the utterance contours of 65 Low-Level Descriptors (LLDs) and their deltas estimated from the speech signal every 10 ms. Table 4 summarizes the LLDs included as frame-level features. The set of statistic functionals applied to the LLD contours at the utterance level includes percentiles, modulations, moments, peaks and regressions, and is presented in Table 5. The LLDs and functionals are described in detail in [27].

In an attempt to preserve the features that are the most relevant to the task at hand and to reduce the complexity of the classification stage, we applied a correlation-based feature subset selection evaluator with a best-first search method [28]. This is a supervised dimensionality reduction technique that evaluates the worth of a subset of features by considering their individual predictive ability along with the degree of redundancy between features. It generally chooses subsets that have low intercorrelation and are highly correlated with the expected classification. We selected the feature subset using the training set and were left with 221 of the original 6015 static features. As the selection procedure resulted in a substantial reduction in the total number of features, we tested classifiers with both the original and the reduced set of features.

3.2 Age Group Classification Experiments

We implemented age group classifiers using linear kernel Support Vector Machines (SVMs) [29, 30] trained with the Sequential Minimal Optimisation (SMO) algorithm [31]. This combination is known to be robust against overfitting when used in tasks with high-dimensional feature spaces and unbalanced class distributions. We normalised feature values to be in the range [0, 1] prior to classifier training, estimating the normalisation parameters on the training set and then applying them to the training, development and evaluation test sets. We investigated SVM optimisation by training models with different values for the complexity parameter "C" of the SMO algorithm and by choosing the one that obtained the highest performance on the development test set. As we had two feature sets, one containing all the features and the other containing the automatically selected subset of features, we trained SVMs using both sets and chose the complexity parameter independently. Fig. 1 represents the classification results on the development and evaluation test sets with different complexity values for models trained with all the features and for models with the automatically selected subset of feature.

3.3 Results and Discussion

To evaluate the age group classifiers, we used the Unweighted Average Recall (UAR) metric. Compared with the standard accuracy metric, this metric allows a more meaningful assessment when evaluating datasets with unbalanced class distributions. For our three-class problem at hand – children (C), young to middle-aged adults (A) and elderly people (E), the UAR metric is calculated as (Recall(C)+Recall(A)+Recall(E))/3. That is, the number of instances per class is intentionally ignored.

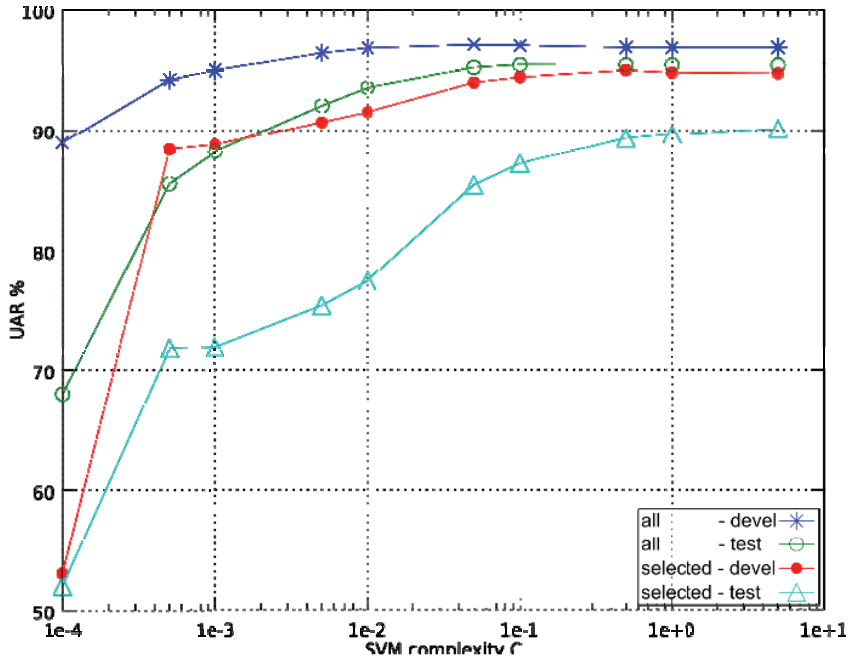

Fig. 1. Age group classification results (UAR %) on the development and evaluation test sets for the SVM models trained with all the features and with the automatically selected subset of features

After training 20 classifier models with different "C" values for the SMO complexity, we chose the model with the highest UAR on the development test set as our age group classifier. Table 6 shows that this model, trained with all the features, obtains a very high performance on the evaluation test set. Table 7 presents the corresponding age group confusion matrix for the evaluation test set.

Table 6. Age group classification results (UAR%) obtained on the development and evaluation test sets

SVM Classifier	C	Devel	Eval
all features	0.05	97.13	95.24
selected features	0.5	95.00	89.33

Table 7. Age group confusion matrix for the evaluation test set. Values in the table indicate the percentage of automatically classified utterances.

	CNG	YMA	EASR
CNG	**98.6**	0.8	0.6
YMA	0.8	**92.0**	7.2
EASR	1.8	3.2	**95.0**

Like most human physiological processes, aging and the consequent vocal aging are gradual processes. The aging of speech is not only affected by chronological age but people's lifestyles (e.g. smoking, alcohol consumption, psychological stress, abuse and overuse of vocal cords) [32], as well. While it might be impossible to determine the exact age from which an individual's speech could be considered elderly, studies usually regard 60-70 years of age as the minimum age range for elderly

speech [10]. For our experiments, we decided to consider 60 years of age as the boundary between young to middle-aged adults and the elderly. Our choice is in line with the choice that was made for selecting speakers for the EASR Corpus (see [18]). However, the "fuzziness" of the age boundary is reflected in the number of erroneously classified utterances from young to middle-aged and elderly speakers. At the other end of the age range, the number of incorrectly classified utterances from children is very small. The performance is undoubtedly boosted by the complete lack of test data from speakers aged 11- 24.

4 ASR Experiments

This section describes the ASR experiments carried out to test the potential benefits of automatic age group classification in ASR. We used three different sets of Hidden Markov Model (HMM) models for the experiments: "standard" AMs trained using young to middle-aged adults' speech (hereafter "SAM"), as well as two separate sets of AMs specifically optimised for children's and elderly people's speech (hereafter "CAM" and "EAM"), respectively. These three sets of AMs are discussed in Section 4.1.

The pronunciation lexicon and language model remained the same across the experiments. The lexicon contained an average of one pronunciation per word, represented using a set of 38 phone labels. For language modelling purposes, we authored a grammar that allowed the words in the phonetically rich sentences of our corpora (see Section 2) to appear 1 to 33 times; the phonetically rich sentences in our data have a minimum of one and a maximum of 33 words. The grammar can be considered as a very simplified language model and, although it yields unrealistically high WERs, the ASR results are comparable across the different evaluation test sets used in our experiments.

4.1 Acoustic Modelling

"Standard" Acoustic Models (SAM). The "standard" AMs originate from the European Portuguese language pack that comes with Microsoft Speech Platform Runtime (Version 11) [33]; a language pack incorporates the language-specific components necessary for ASR: AMs, a pronunciation lexicon, and grammars or a language model. The AMs comprise a mix of gender-dependent whole-word models and cross-word triphones trained using several hundred hours of read and spontaneous speech collected from young to middle-aged adult speakers. In other words, children and the elderly fall outside of this target demographic. The models also include a silence model, a hesitation model for modelling filled pauses, and a noise model for modelling human and non-human noises. More detailed information about the "standard" AMs is not publicly available, it being commercially sensitive information.

Acoustic Models Optimised for Children's Speech (CAM). The "standard" AMs were optimised for children's speech by retraining the female AMs with the training set extracted from the CNG Corpus (see Table 1), regardless of the children's gender. The motivation for only retraining the female AMs with children's speech stems from the fact that the acoustic characteristics of children's speech are more similar to adult female speech than to adult male speech [1, 2]. The hesitation and noise models of the baseline recogniser were retrained utilising the annotations for filled pauses and

noises that are available in the corpus (see Section 2). The children's speech models are discussed in more detail in [12].

Acoustic Models Optimised for Elderly Speech (EAM). The "standard" AMs were optimised for elderly speech by retraining the male and female AMs with the male and female data in the training set extracted from the EASR Corpus (see Table 1), respectively. The hesitation and noise models of the baseline recogniser were again retrained utilising the annotations available in the corpus (see Section 2).

4.2 Experimental Set-Up

We carried out ASR experiments on the children's, young to middle-aged adults', and elderly people's speech in our test sets using three different set-ups: 1) Speech recognised using the "standard" AMs (SAM) regardless of the age group of the speaker, 2) Speech recognised using the AMs corresponding to the known age group of the speaker (CAM, SAM or EAM), and 3) Speech recognised using the AMs corresponding to the automatically determined age group of the speaker (AM-Auto). The results of set-up 1) represent a situation in which we have no way of knowing the speaker's age and must use the "standard" AMs. These results represent our baseline; should no alternative, age group –specific AMs and ways of selecting the correct set of age group –specific AMs be available, we would have to use the "standard" AMs. The results of set-up 2) represent the best achievable results ("oracle") because we are using AMs that have been selected using the known age groups of the speakers. The results of set-up 3) represent the results achieved using the automatic age group classifier described in Section 3.

4.3 Results and Discussion

Table 8 illustrates the ASR results for the experimental set-ups described in Section 4.3. Because of the simplified language model that we used (see Section 4), the overall WERs are high. However, the results show that recognising children's and elderly people's speech using AMs optimised for their own age group can lead to significant improvements in ASR performance over the baseline performance. The results also show that, although the ASR results achieved using the automatic age group classifier are worse than the best achievable ASR results ("*oracle*") in the case of children's and elderly speech, as could be expected, the delta is very small. The ASR results achieved using the automatic age group classifier on young to middle-aged adults' speech are slightly better than the ASR results achieved using the baseline recogniser. This means that, for some of those speakers' voices, the age group –specific models were acoustically a better match than the default models. We intend to analyse if this is, for instance, related to them being close to the age group boundary that we selected.

Table 8. Automatic speech recognition results (WERs).

Eval Set	SAM	CAM	EAM	AM-Auto
CNG	78.3%	46.1%		47.8%
YMA	56.4%			56.3%
EASR	55.9%		48.2%	48.9%

5 Conclusions and Future Work

This paper presented an age group classification system that automatically determines the age group of a speaker from an input speech signal. There were three possible age groups: children, young to middle-aged adults and the elderly. What sets our study apart from other studies on age classification is that we used our age group classifier together with an automatic speech recogniser. More specifically, we carried out ASR experiments in which the automatically determined age group of speakers was used to select age group –specific acoustic models, i.e., acoustic models optimised for children's, young to middle-aged adults' and elderly people's speech. The ASR results showed that using the results of the age group classifier to select age group –specific acoustic models for children and the elderly leads to considerable gains in automatic speech recognition performance, as compared with using "standard" acoustic models trained with young to middle-aged adults' speech for recognising their speech, as well. This finding can be used to improve the speech recognition performance of speech-enabled applications that are used by people of widely varying ages. What makes the approach particularly interesting is that it is a user-friendly alternative for speaker adaptation, which requires the user to spend time training the system.

Both children and elderly people are more likely to interact with computers using everyday language and their own commands, even when a specific syntax is required [5-8]. In future research, we will attempt building age group –specific language models and lexica, and select the optimal language model and lexicon using the results of our age group classifier. We hypothesise that such an approach could lead to further gains in ASR performance.

One of the limitations of this study is that we did not have any acoustic model training data or test data from 11-24-year-old speakers. This probably led to unrealistically good age group classification performance in the case of children and young to middle-aged adults. Therefore, in future research, we also intend to record test data from 11-24-year-old European Portuguese speakers, optimise acoustic models for representatives of that age group, and rerun age group classification and ASR experiments. We expect this to be a challenging age group to work with, as the values of children's acoustic parameters converge to adult levels at around 13-15 years of age [1].

Acknowledgements. This work was partially supported by: (1) the QREN 5329 Fala Global project, which is co-funded by Microsoft and the European Structural Funds for Portugal (FEDER) through POR Lisboa (Regional Operational Programme of Lisbon), as part of the National Strategic Reference Framework (QREN), the national program of incentives for Portuguese businesses and industry; (2) the EU-IST FP7 project SpeDial under contract 611396 and (3) Fundação para a Ciência e a Tecnologia, through project PEst-OE/EEI/LA0008/2013.

References

[1] Lee, S., Potamianos, A., Narayanan, S.: Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters. J. Acoust. Soc. Am. 10, 1455–1468 (1999)

[2] Huber, J.E., Stathopoulos, E.T., Curione, G.M., Ash, T.A., Johnson, K.: Formants of Children, Women and Men: The Effects of Vocal Intensity Variation. J. Acoust. Soc. Am. 106(3), 1532–1542 (1999)

[3] Xue, S., Hao, G.: Changes in the Human Vocal Tract Due to Aging and the Acoustic Correlates of Speech Production: A Pilot Study. Journal of Speech, Language, and Hearing Research 46, 689–701 (2003)
[4] Pellegrini, T., Hämäläinen, A., Boula de Mareüil, P., Tjalve, M., Trancoso, I., Candeias, S., Sales Dias, M., Braga, D.: A Corpus-Based Study of Elderly and Young Speakers of European Portuguese: Acoustic Correlates and Their Impact on Speech Recognition Performance. In: Interspeech, Lyon (2013)
[5] Narayanan, S., Potamianos, A.: Creating Conversational Interfaces for Children. IEEE Speech Audio Process. 10(2), 65–78 (2002)
[6] Strommen, E.F., Frome, F.S.: Talking Back to Big Bird: Preschool Users and a Simple Speech Recognition System. Educ. Technol. Res. Dev. 41(1), 5–16 (1993)
[7] Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R.: Recognition of Elderly Speech and Voice-Driven Document Retrieval. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, AZ, pp. 145–148 (1999)
[8] Takahashi, S., Morimoto, T., Maeda, S., Tsuruta, N.: Dialogue Experiment for Elderly People in Home Health Care System. In: Matoušek, V., Mautner, P. (eds.) TSD 2003. LNCS (LNAI), vol. 2807, pp. 418–423. Springer, Heidelberg (2003)
[9] Teixeira, V., Pires, C., Pinto, F., Freitas, J., Dias, M.S., Mendes Rodrigues, E.: Towards Elderly Social Integration using a Multimodal Human-computer Interface. In: Proc. of the 2nd International Living Usability Lab Workshop on AAL Latest Solutions, Trends and Applications, AAL 2012, Milan (2012)
[10] Wilpon, J.G., Jacobsen, C.N.: A Study of Speech Recognition for Children and Elderly. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, GA, pp. 349–352 (1996)
[11] Potamianos, A., Narayanan, S.: Robust Recognition of Children's Speech. IEEE Speech Audio Process 11(6), 603–615 (2003)
[12] Hämäläinen, A., Miguel Pinto, F., Rodrigues, S., Júdice, A., Morgado Silva, S., Calado, A., Sales Dias, M.: A Multimodal Educational Game for 3-10-year-old Children: Collecting and Automatically Recognising European Portuguese Children's Speech. In: Workshop on Speech and Language Technology in Education, Grenoble (2013)
[13] Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Sales Dias, M., Braga, D.: Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese. In: IberSPEECH, Madrid (2012)
[14] Vipperla, R., Renals, S., Frankel, J.: Longitudinal Study of ASR Performance on Ageing Voices. In: Interspeech, Brisbane, pp. 2550–2553 (2008)
[15] Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., Wong, M.: The PF_STAR Children's Speech Corpus. In: Interspeech, Lisbon (2005)
[16] Hämäläinen, A., Rodrigues, S., Júdice, A., Silva, S.M., Calado, A., Pinto, F.M., Dias, M.S.: The CNG Corpus of European Portuguese Children's Speech. In: Habernal, I. (ed.) TSD 2013. LNCS (LNAI), vol. 8082, pp. 544–551. Springer, Heidelberg (2013)
[17] Cucchiarini, C., Van Hamme, H., van Herwijnen, O., Smits, F.: JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality. In: Language Resources and Evaluation, Genoa (2006)
[18] Hämäläinen, A., Pinto, F., Sales Dias, M., Júdice, A., Freitas, J., Pires, C., Teixeira, V., Calado, A., Braga, D.: The First European Portuguese Elderly Speech Corpus. In: IberSPEECH, Madrid (2012)

[19] Hämäläinen, A., Avelar, J., Rodrigues, S., Sales Dias, M., Kolesiński, A., Fegyó, T., Nemeth, G., Csobánka, P., Lan Hing Ting, K., Hewson, D.: The EASR Corpora of European Portuguese, French, Hungarian and Polish Elderly Speech. In: Langauge Resources and Evaluation, Reykjavik (2014)
[20] Minematsu, N., Sekiguchi, M., Hirose, K.: Automatic Estimation of One's Age with His/Her Speech Basedupon Acoustic Modeling Techniques of Speakers. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, pp. 137–140 (2002)
[21] Dobry, G., Hecht, R., Avigal, M., Zigel, Y.: Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal. IEEE Transactions on Audio, Speech & Language Processing 19(7), 1975–1985 (2011)
[22] Bahari, M., McLaren, M., Van Hamme, H., Van Leeuwen, D.: Age Estimation from Telephone Speech Using i-Vectors. In: Interspeech, Portland, OR (2012)
[23] Neto, J., Martins, C., Meinedo, H., Almeida, L.: The Design of a Large Vocabulary Speech Corpus for Portuguese. In: European Conference on Speech Technology, Rhodes (1997)
[24] Eyben, F., Wollmer, M., Schuller, B.: openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: ACM International Conference on Multimedia, Florence, pp. 1459–1462 (2010)
[25] Meinedo, H., Trancoso, I.: Age and Gender Detection in the I-DASH Project. ACM Trans. Speech Lang. Process. 7(4), 13 (2011)
[26] Schuller, B., Steidl, S., Batliner, A., Noeth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B.: The Interspeech 2012 Speaker Trait Challenge. In: Interspeech 2012, Portland, OR (2012)
[27] Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M., Scherer, K.R.: On the Acoustics of Emotion in Audio: What Speech, Music and Sound Have in Common. Frontiers in Psychology, Emotion Science, Special Issue on Expression of Emotion in Music and Vocal Communication 4(Article ID 292), 1–12 (2013)
[28] Hall, M.: Correlation-Based Feature Subset Selection for Machine Learning. Hamilton, New Zealand (1998)
[29] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11 (2009)
[30] Platt, J.: Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning (1998)
[31] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13(3), 637–649 (2001)
[32] Linville, S.E.: Vocal Aging. Singular, San Diego (2001)
[33] Microsoft Speech Platform Runtime (Version 11), `http://www.microsoft.com/en-us/download/details.aspx?id=27225` (accessed March 25, 2013)