INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Infected Phonemes: How a Cold Impairs Speech on a Phonetic Level

Johannes Wagner[1], *Thiago Fraga-Silva*[2], *Yvan Josse*[2],
Dominik Schiller[1], *Andreas Seiderer*[1], *Elisabeth André*[1]

[1]Human Centered Multimedia, Augsburg University, Augsburg, Germany
[2]Vocapia Research, 28 rue Jean Rostand, 91400 Orsay, France
{wagner,schiller,seiderer,andre}@hcm-lab.de, {thfraga,josse}@vocapia.com

Abstract

The realization of language through vocal sounds involves a complex interplay between the lungs, the vocal cords, and a series of resonant chambers (e.g. mouth and nasal cavities). Due to their connection to the outside world, these body parts are popular spots for viruses and bacteria to enter the human organism. Affected people may suffer from an upper respiratory tract infection (URTIC) and consequently their voice often sounds breathy, raspy or sniffly. In this paper, we investigate the audible effects of a cold on a phonetic level. Results on a German corpus show that the articulation of consonants is more impaired than that of vowels. Surprisingly, nasal sounds do not follow this trend in our experiments. We finally try to predict a speaker's health condition by fusing decisions we derive from single phonemes. The presented work is part of the INTERSPEECH 2017 Computational Paralinguistics Challenge.
Index Terms: speech recognition, computational paralinguistics, health condition

1. Introduction

Speech conveys a manifold of information, which extends way beyond the linguistic content. This additional information is hidden in the way *how* something is said rather than *what* is being said. The field of computational paralinguistics deals with the automatic extraction and analysis of the *non-verbal* aspects of speech. Recent work in the field has focused on various tasks such as the recognition of emotions [1], depression and suicidal tendency [2], laughter [3] or conflicts [4]. In this paper we aim at assessing a speaker's health by automatically distinguishing between speech under cold and speech under "normal" circumstances. The task is proposed as part of the INTERSPEECH 2017 Computational Paralinguistics Challenge [5].

People suffering from a common cold may show a combination of various symptoms such as nasal obstruction and stuffiness, hoarseness, coughing, or sneezing [6]. These symptoms directly affect the vocal tract causing a temporary speech disorder. Such speaker-related variability are one of the main difficulties in speech signal processing [7]. Yet, due to the lack of adequate speech databases, research on cold speech is still rare. Therefore, the corpus provided for the Cold sub-challenge containing audio recordings of 630 German speakers with a total duration of approximately 45 hours offers an excellent "playground" to study how a cold influences a speaker's voice. The gained knowledge can help improving the robustness of speech related tasks such as speech recognition, speaker identification [7] or emotion recognition [8].

The detection system we present starts from the assumption that the symptoms of an upper respiratory tract infection (URTIC) are not equally audible throughout an utterance. This is reasonable since place and manner of articulation differs for different phonemes. For instance, we would expect that nasal sounds like /n/ and /m/, which are produced with a lowered velum allowing air to escape freely through the nose, are especially affected by a stuffy nose. Hoarseness, on the other hand, limits the ability of the vocal cords to vibrate and we would reckon a particularly strong effect on the production of vowels such as /a/ and /e/. Taken this into account, we decided to investigate the data on a phonetic level. In particular, we aim at answering the following questions:

1. How does a cold affect phone articulation?
2. Are certain phonemes or phoneme classes especially affected by a cold?
3. Can we predict the health state of a speaker by fusing decisions derived on a phonetic level?

2. Related Work

In the 1990s, Renetta G. Tull and colleagues investigated how a cold influences certain speech features. For instance, they found noticeable patterns in the lower-order mel-cepstral coefficients [9] and measured more noisy portions in cold speech caused by hoarseness and coughing [10]. Phonetic transcriptions of cold and healthy sessions revealed changes in place of articulation and that pauses and epenthetic syllables are not constant throughout all sessions [10]. A detailed analysis of the vowels /i/, /a/, /æ/ showed that the formats F1 and F2 are lowered for the cold condition [11]. In a more recent work by Philip Rose [12] the author reports long-term F0 distribution obtained in good health and when suffering from a severe laryngitis. In the latter case, he measured a significantly lower mean and standard deviation.

Approaching speech on a phonetic level cannot just provide a better understanding how phone articulation is affected by external influences, but has proven useful in paralinguistic classification tasks, too. Lee and colleagues [13] used a segmentation into five phoneme classes – vowel, stop, glide, nasal and fricatives – to investigate the effects of emotions on the different speech sounds. In their experiments, they trained separate Hidden Markov Models (HMMs) for each phoneme class based on short-term spectral features. They found that the phoneme-based classification system achieved significant better results than their baseline classifiers. This leads to the conclusion that emotions have a stronger effect on the articulation of certain phonemes.

Similar observations have been reported by Schuller et al. [14]. In their experiments, they trained multiple emotion classification models on phoneme and word level segments. Both approaches outperformed common general models when provided enough training material for each unit. Those findings confirm the effectiveness of specialized, segment-based classi-

http://dx.doi.org/10.21437/Interspeech.2017-1066

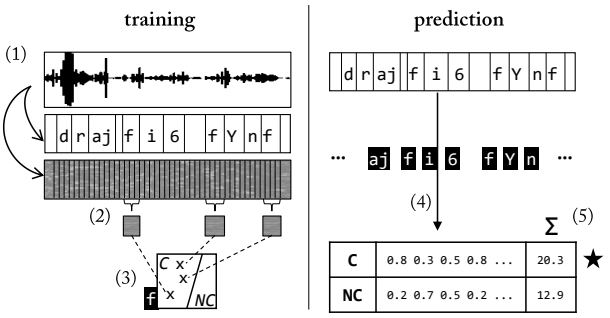

Figure 1: *Phonetic-based detection system. Training: (1) For all utterances in the training set we derive phonetic transcriptions and low-level features. (2) According to the start and end positions of the phonemes we cut out feature chunks and average them. (3) For each phoneme we train a linear model to discriminate cold (C) vs. non-cold (NC). Pre diction: (4) We collect individual predictions for the phonemes of an utterance and fuse them, e.g. by sum rule (5).*

fication systems in order to recognize variances in the manner of speaking.

The fact that recognition accuracy in paralinguistic recognition tasks can be improved by switching to shorter units leads to the assumption that meaningful information may not be equally distributed across an utterance. Instead it may be locally concentrated in certain "hot" spots. Locating such "hot" spots could help building more coherent models. In an earlier work we have investigated this within the context of personality trait detection [15]. We proposed a cluster-based approach, which aims at identifying frames that will likely carry cues about the personality. Afterwards, we pruned the speech utterances and kept only most promising frames. In this paper, we follow up the idea, but instead of working on a frame level, we investigate a possibly more coherent unit: phonemes.

3. Methodology

Figure 1 gives an overview of the phoneme-based detection system. It starts with a phonetic transcription we derive from an automatic speech recognizer (ASR) described in Section 3.1. Based on the transcriptions we train classification models for single phoneme classes (Section 3.3). To train the models we compare different types of low-level features (LLF), which we introduce in Section 3.2. Finally, to predict cold speech on utterance level, we combine individual scores as described in Section 3.4.

3.1. Phoneme Detection

The phonetic transcription is obtained using a large vocabulary continuous speech recognizer (LVCSR), trained on broadcast speech in German. The *acoustic model* is based on deep neural networks (DNNs). The DNN is trained on 150 hours of speech collected during the Quaero project and has about 10M parameters and 4 hidden layers. The softmax output layer targets about 10k tied-states of hidden Markov phone models. Input is based on perceptual linear prediction (PLP) features. Speaker adaptive training (SAT) with constrained maximum likelihood linear regression (CMLLR) [16] is used to estimate the model parameters. *Language models* are backoff n-gram models build with a 2 giga-word corpus collected by Vocapia (newspapers archives, web news, etc.) and estimated with Kneser-Ney smoothing [17]. The *pronunciation lexicon* is partially obtained using a data-driven grapheme-to-phoneme converter [18]. The phone set comprises 46 phonemes and has three special units to model silence, breath and filler words.

Recognition is carried out on a single pass, which generates a lattice containing word and phonetic information. Viterbi decoding is applied to obtain the best word and phonetic sequences for each utterance. An alternative to generate phoneme sequences is to use a purely phonetic decoder. However, better phoneme error rates were obtained using the LVCSR system.

3.2. Low-level Features

MFCC: Mel-frequency cepstral coefficients (MFCCs) are well known for their application in ASR systems, and have shown good results in emotional speech recognition tasks, too [19]. In this work we use the implementation provided by the OpenSMILE toolbox[1] [20] and extract 13 coefficient (including the 0^{th}) using a sliding window of 25 ms at a frame step of 10 ms (*mfcc*). In addition, we calculate a second set, which models the temporal flow by adding 1^{st} and 2^{nd} derivatives (*mfccd*).

IIF: To reduce the influence of speaker-related variabilities one can apply speaker-adaption techniques, e.g. by extracting features that compensate for different vocal tract lengths [21]. Müller et al. [22] propose the use of contextual Invariant-Integration Features (IIF). The features are designed to be invariant to translations along the subband-index space of the time-frequency representation. In this work we adopt code generously provided by the authors[2] and again obtain the features at a frame step of 10 ms (*iif*).

CMLLR: Inter-speaker variability can be dealt with adaptive training methods [23]. The goal is to project the acoustic features into a canonical feature space, common to all speakers. A CMLLR (constrained maximum likelihood linear regression) transform [16] is used for feature projection during training and recognition phases. Here, the adaptive features (*cmllr*) are generated as follows. First, a 14-dimensional feature vector containing energy, pitch and 12 MFCCs is generated every 10 ms. Mean and variance normalization is applied, and first and second derivatives are calculated to form a 42-dimensional vector. Then, 9 vectors (4 on the left context and 4 on the right) are concatenated. Linear discriminant analysis is applied to reduce the feature vector to 40 dimensions. Finally, a CMLLR transform is applied.

3.3. Linear Classification

As classification model we use a linear Support Vector Machine (SVM) provided by LIBLINEAR[3] – a Library for Large Linear Classification [24]. Since the implementation does not use kernels, training time is significantly reduced even for large input sets composed of several ten thousand samples. We use grid search to optimize the solver (0=L2-regularized logistic regression, 3=L2-regularized L1-loss, 5=1-regularized L2-loss), the complexity (1, 1e-1, 1e-2, 1e-3, 1e-4), the learning rate (1, 1e-1, 1e-2, 1e-3), and the optional bias term (-1=none 1, 0, 1e-1).

[1] http://audeering.com/technology/opensmile/

[2] https://www.isip.uni-luebeck.de/downloads/computeiif-matlab.html – our tests base on 'iifset30' and use 'LDA20' as reduction matrix followed by MLLT transformation.

[3] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

Table 1: *Original alphabet A and a slightly modified version A*, as well as, the two groups C_6 and C_3. Absolute frequencies on the whole corpus on top of cells (×1000).*

	74	73	40	40	41	3.7	46	72	0.2	22	24	3	83	17	37	8	6	8	43	19	3.8	5	13
A	a	A	@	e	E	E:	i	I	1	o	O	2	oe	u	U	y	Y	j	aj	aw	Oj	6	6_^
*A**	a	A	@	e	E*		i*			o	O	oe*		u*		y*		j	aj	aw	Oj	6*	
C_6 / C_3	*vowel*																	*glide*					
	28	78	32	29	11	155	34	41	19	94	28	42	17	44	0.5	61	0.5	110	175	35	10	52	1.6
A	b	d	g	k	p	t	C	f	h	s	S	v	x	z	Z	l	l_=	r	n	n_=	N	m	m_=
*A**	b	d	g	k	p	t	C	f	h	s	S	v	x	z*		l*		r	n*		N	m*	
C_6	*stop*						*fricative*									*liquid*			*nasal*				
C_3	*consonant*																						

We balance the number of samples per class by randomly duplicating samples of classes that are underrepresented. Finally, features values are scaled between -1 and 1.

3.4. Fusion

To come up with a single label for the whole utterance, we sum up probabilities for the phonemes in the utterance and decide in favour of the class with the higher score (sum rule). That is, we retrieve probabilities by consulting the models trained for individual phonemes (or enclosing groups). Each classification provides confidence that a particular phoneme within the utterance was pronounced under cold or not.

4. Results

We now report results from several experiments we performed with the detection system described in the last section. Performance will be given with respect to the Unweighted Average Recall (UAR), the official evaluation measurement of the challenge. The class C denotes samples from users who suffered from a cold, while the class NC denotes healthy examples. We stick to the training and development set proposed by the challenge organizers with each set consisting of roughly 10k utterances lasting between 3 s and 10 s (for details please see [5]).

4.1. Grouping

First, we reduced the original phone set from 46 to 35 phonemes (henceforth *A**) by merging similar phones or rare phones together. *A** was further grouped into vowels, stops, fricatives, liquids, nasal sounds, and a class denoted as *glide* combining semi-vowels and gliding vowels (diphthongs) [25]. A second even coarser grouping differentiates vowels, consonants and glides. Table 1 gives an overview of the original and modified alphabet using the X-SAMPA notation, as well as, the two coarser groups C_6 and C_3.

4.2. Phoneme Level

Table 2 lists the performance on phoneme level. For each phoneme class a model is trained on the C vs. NC condition and performance is evaluated on the development set (Section 3.3). Results are shown for the *iif* feature set, which gave the highest performance. To improve readability entries are sorted by UAR and the table is split into two columns. Phonemes in the left column generally outperform those in the right column.

For the modified alphabet *A** scores are within a range of more than 10 %. The ranking is headed by the vowel /@/ yielding a 65.2 % UAR. However, except for /o/ and /aw/ vowels and glides are only found in the right column. Surprisingly, this also counts for *nasal* sounds. Most *stop*, *fricative* and *liquid* sounds, on the other hand, occur in the left column. This is also reflected by the results we gain for the coarser classes listed at the bottom of the table. No correlation between phoneme frequency and classification performance can be observed.

Table 2: *Results on phoneme level ranked by UAR measured on the development set with iif features. f: Relative frequency. UAR: Unweighted Average Recall.*

	$f^{\%}$	UAR$^{\%}$		$f^{\%}$	UAR$^{\%}$
@	2.3	**65.2**	O	1.4	61.4
z*	2.5	63.8	u*	3.1	61.3
p	0.6	63.7	v	2.4	61.3
x	1.0	63.6	l*	3.6	61.2
o	1.3	63.5	A	2.2	61.1
aw	1.1	63.0	i*	6.8	60.9
C	2.0	62.5	a	4.2	60.8
r	6.3	62.4	oe*	5.0	60.3
b	1.6	62.4	e	2.3	60.3
k	1.7	62.3	m*	3.1	60.2
s	5.4	62.3	j	0.5	60.0
t	8.9	62.2	E*	2.6	59.4
S	1.6	62.1	h	1.1	58.2
f	2.4	61.9	N	0.6	58.1
d	4.5	61.9	aj	2.5	57.9
g	1.9	61.8	6*	1.1	56.3
n*	12.0	61.7	y*	0.8	55.0
			Oj	0.2	54.5
consonant	63.0	**62.7**	*vowel*	31.7	60.8
liquid	9.8	62.3	*glide*	5.3	60.3
stops	19.1	62.2	*nasal*	15.7	59.6
fricative	18.3	62.2			

4.3. Utterance Level

To measure performance on utterance level, we fuse individual decisions we receive for the contained phonemes (Section 3.4). Table 3 summarizes results for different phoneme groups and feature sets (Section 3.2).

We observe the highest score for the *iif* feature set with *A** yielding a 67.6 % UAR. The performance of C_6 and C_3 is generally below that of *A**. No improvement is gained by adding deltas (see *mfcc* vs. *mfccd*). For *cmllr* – despite a low UAR – we observe the highest NC recall of all sets (70.8 %). On the other hand, *iif* has the highest recall for C (72.9 %). Merging the two sets into the superset *iif+cmllr* improves results for C_6 and C_3 but not *A**. On the development set results are much in line with the baseline reported by the challenge organizers (66.1 % in the best case). On the test set, however, our approach only reached an UAR of 63.6% and hence clearly stayed behind the official baseline of 71.0% [5].

Table 3: *Performance on utterance level for different feature sets on the development set. #: feature dimension. C: cold. NC: non-cold. UAR: Unweighted Average Recall.*

		$A*$			C_6	C_3
	#	C	NC	UAR$^{\%}$	UAR$^{\%}$	UAR$^{\%}$
iif	60	72.9	62.3	**67.6**	65.9	66.0
mfcc	13	67.6	63.0	65.3	64.2	63.8
mfccd	39	67.9	62.5	65.2	63.7	65.0
cmllr	40	50.0	70.8	60.4	58.9	58.9
iff+cmllr	100	71.1	63.8	67.5	**66.3**	**67.1**

5. Discussion

Before we present our interpretation of the results from the last section, it is worthwhile to have a look at the distributions we measured for some common speech features as they allow us to gain insights how a cold affects the articulation of certain phonemes. In fact, our observations verify on a large volume of data the findings by Tull [10, 9, 11] and Rose [12], which were based on a small number of subjects only. As can be seen in Figure 2 the pitch of all vowels is lowered during a cold. This can be explained with the fact that a swollen vocal cord vibrates more slowly, so that the voice sounds lower than it usually is. For stops and liquids the Signal-to-Noise (S/N) ratio decreases, which indicates an increase in noise – probably caused by hoarseness and coughing. Finally, for most fricatives the first two mel-cepstral coefficients are lower for cold speech. This suggests a change in timbre, for instance due to a hoarse voice.

Regarding the question whether certain phonemes are more affected by a cold, the fact that in our experiments the performance of individual phonemes varies by more than 10 %, strongly supports this assumption (Section 4.2). Looking at broader trends, we noticed that in our data consonants outperform vowels and glides. A possible reason could be that vowels and glides are produced with a relatively open vocal tract through which air flows with little resistance [26]. Hence, if some parts of the vocal tract are slightly swollen this will not immediately have an audible effect. It may, however, affect consonants, which already involve some degree of obstruction of the airflow. In fact, we believe that the provided corpus contains few prototypical examples of a cold. Listening to the examples we often found it hard to judge if a speaker suffers from a cold or not. This also explains a rather moderate baseline of 71 % UAR at a 50 % chance level [5]. With that said, we may also explain another rather surprising finding. Intuitively we would expect that a cold is especially audible with the nasal sounds. This seems natural since a stuffy nose – a common symptom of a cold – blocks the airflow through the oral cavity causing hypernasality. However, although consonants generally performed well, nasal sounds did not follow this trend. This again can be taken as a sign that the speakers in the provided data show a wide range of rather subtle symptoms. In the end, we may experience what we already know from other recognition tasks: prototypical behavior in real-life data remains the exception [27].

When we tried to fuse decisions of single phonemes to predict the health condition on utterance level, results remained behind the official challenge baseline (Section 4.3). Hence, we can conclude if one is only interested in detecting cold speech, processing on a phonetic level may not be worth the effort.

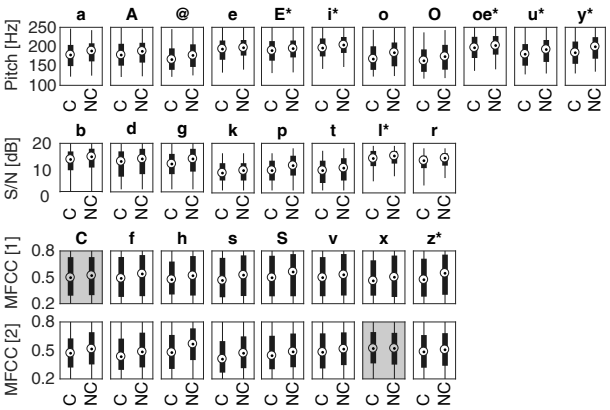

Figure 2: *Box plots of feature distributions for different phonemes in C vs. NC condition (calculated over combined train and devel set). Plots that are not significantly different are marked by a gray background ($p < 0.05$).*

6. Conclusion

In the presented work, we have investigated methods to detect cold speech using a phonetic-based approach on a large corpus of German. Based on phonetic transcriptions we trained models to evaluate how a cold impairs speech on a phonetic level.

The corpus for our experiments consists of approximately 45 hours of speech and was provided by the organizers of the INTERSPEECH 2017 Computational Paralinguistics Challenge [5]. This considerable volume allowed us to verify findings from earlier studies, which were based on a small number of subjects only. By observing distributions of common speech features, we proved that a cold could lead to a lowered pitch, introduce additional noise and change the timbre of the voice (causing a decrease in the lower mel-cepstral coefficients).

We also made findings not reported in literature. In our experiments the articulation of consonants seemed to be more impaired by a cold than that of vowels. A possible (yet hypothetical) explanation could be that during the production of consonants the airflow is obstructed to some degree, so that an audible effect is already notable even if the vocal tract is only slightly swollen. Nasal sounds, however, did not follow this trend. We explain this with the fact that the examined data contains mainly non-prototypical examples of cold speech. In fact, speakers with a completely stuffed nose are rather the exception in the data.

Features reducing the influence of speaker-related variabilities by compensating for different vocal tract length showed better results compared to standard MFCC features in our experiments. However, when fusing decisions to predict the health condition of longer utterances, results remained behind the official challenge baseline. Hence, in future work our approach might be improved by considering co-articulatory effects. Since its neighbors always influence the articulation of a phoneme, it might be beneficial to incorporate a larger context, for instance, by considering bi- or triphone combinations.

7. Acknowledgements

The work has received funding from the European Commission under the contract number H2020-RIA-645012, KRISTINA, and the European Unions Horizon 2020 research and innovation program under grant agreement No 645378, ARIA-VALUSPA.

8. References

[1] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[2] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[3] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran, "An event driven fusion approach for enjoyment recognition in real-time," in *International Conference on Multimedia (MM)*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 377–386.

[4] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, "Random discriminative projection based feature selection with application to conflict recognition," *IEEE Signal Processing Letters*, vol. 22, no. 6, pp. 671–675, 2015.

[5] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, A. S. Warlaumont, G. Hidalgo, S. Schnieder, C. Heiser, W. Hohenhorst, M. Herzog, M. Schmitt, K. Qian, Y. Zhang, G. Trigeorgis, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *INTERSPEECH 2017, Conference of the International Speech Communication Association (ISCA), Stockholm, Sweden*, 2017.

[6] D. A. Tyrrell, S. Cohen, and J. E. Schlarb, "Signs and symptoms in common colds." *Epidemiol Infect*, vol. 111, no. 1, pp. 143–156, Aug 1993.

[7] T. F. Zheng, Q. Jin, L. Li, J. Wang, and F. Bie, "An overview of robustness related issues in speaker recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–10.

[8] J. Krajewski, E. Nth, and A. Batliner, *Der Mensch im Mittelpunkt technischer Systeme*, ser. Mensch-Maschine-Interaktion. VDI-Verlag, 2009, vol. 22, no. ISBN 978-3-18-302922-8, pp. 98–103.

[9] R. G. Tull, J. C. Rutledge, and C. R. Larson, "Cepstral analysis of "cold-speech" for speaker recognition: A second look," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2760–2760, 1996.

[10] R. G. Tull and J. C. Rutledge, "Analysis of "cold-affected" speech for inclusion in speaker recognition systems." *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2549–2574, 1996.

[11] R. G. Tull, *Advances in Phonetics: Proceedings of the International Phonetic Sciences Conference (IPS), Bellingham, WA, June 27-30, 1998*, ser. Hermes: Zeitschrift Fur Klassische Philologie. Einzelschrift. F. Steiner, 1999, ch. Returning to format frequencies analysis: a step toward understanding performance problems of cold-speech in automatic speaker recognition systems.

[12] P. Rose, *Forensic Speaker Identification*, ser. Forensic Science, J. Robertson, Ed. Taylor and Francis, 2003, no. ISBN 0-415-27182-7.

[13] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," in *International Conference on Spoken Language Processing (ICSLP)*, 2004, pp. 889–892.

[14] B. Schuller, B. Vlasenko, D. Arsic, G. Rigoll, and A. Wendemuth, "Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition," in *Multimedia and Expo, 2008 IEEE International Conference on*. IEEE, 2008, pp. 1333–1336.

[15] J. Wagner, F. Lingenfelser, and E. André, "A frame pruning approach for paralinguistic recognition tasks," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2012.

[16] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech and language*, vol. 12, no. 2, pp. 75–98, 1998.

[17] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1995, pp. 181–184.

[18] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," *Speech communication*, vol. 50, no. 5, pp. 434–451, 2008.

[19] B. W. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association (ISCA), Antwerp, Belgium*, 2007, pp. 2253–2256.

[20] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in OpenSMILE, the munich open-source multimedia feature extractor," in *International Conference on Multimedia (MM)*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838.

[21] M. Benzeghiba, R. de Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[22] F. Müller, "Invariant features and enhanced speaker normalization for automatic speech recognition," Ph.D. dissertation, University of Lübeck, 2013.

[23] M. J. Gales, "Cluster adaptive training for speech recognition," in *Proc. International Conference on Spoken Language (ICSLP)*, vol. 1998, 1998, pp. 1783–1786.

[24] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jun. 2008.

[25] E. Wilkinson, M. U. D. of German Studies, and S. Studies, *An Introduction to Linguistics for Australian Students of German*. Department of German Studies and Slavic Studies, Monash University, 1998.

[26] V. Dellwo, M. Huckvale, and M. Ashby, *How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 1–20.

[27] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9-10, pp. 1062–1087, Nov. 2011.