Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Constrained Subword Units for Speaker Recognition

Doris Baum[1], *Daniel Schneider*[1], *Timo Mertens*[2], *Joachim Köhler*[1]

[1]Fraunhofer IAIS
Department NetMedia
St. Augustin, Germany

[2]Norwegian University of Science and Technology
Signal Processing Group
Trondheim, Norway

Abstract

Phonetic features have been proposed to overcome performance degradation in spectral speaker recognition in difficult acoustic conditions. The harmful effect of those conditions, however, is not restricted to spectral systems but also affects the performance of the open-loop phone recognisers on which phonetic systems are based. In automatic speech recognition, larger subword units and the use of additional constraints from language models have been employed to improve robustness against adverse acoustic conditions. This paper evaluates the performance of more constrained phone recognition and different subword units for speaker recognition on heterogeneous broadcast data from German parliamentary speeches.

Using phone clusters and a strong language model instead of phones obtained from unconstrained recognition improves the equal error rate from 14.3% to 8.6% on the given data.

1. Introduction

Spectral-based speaker recognition systems, like [1], suffer from performance degradation in the presence of spectral noise or varying spectral characteristics in training and test data. This is a drawback for speaker recognition applications on heterogeneous data, such as speaker verification via telephone or automatic metadata extraction in broadcast multimedia archives.

Especially in the last decade, various approaches have been proposed to overcome this limitation by adding high-level speech information to the recognition process [2]. For example, different features which capture speaker-specific pronunciation variation have been successfully used for speaker recognition tasks: phone n-grams produced by an open-loop phone recogniser [3, 4], time-aligned phone pairs from multiple language phone recognisers [5], and pairs of intended phonemes and their actual phone realisations [6].

However, as the phone recognisers used for pronunciation-feature extraction use spectral features as their input, the same problems that affect spectral-based speaker recognition will likely also have an adverse effect on pronunciation-based systems. Automatic speech recognition (ASR) output has been used in the past to improve spectral and prosodic speaker recognition systems, for example in [7, 8]. In ASR, a common approach to mitigate difficult acoustic conditions is to make use of additional context information during decoding. A lexicon can be employed to recognise larger subword units and using a language model (LM) further constrains the decoding process. In order to apply these methods to speaker recognition, the normalising effect of additional context should ideally be strong enough to compensate for training/test mismatch and noise but not so strong as to remove speaker-specific pronunciation variations. Subword units larger than phones – such as syllables [9] and position-specific phone clusters [10, 11] – appear to capture an appropriate amount of context and thereby provide a suitable granularity of normalisation. In comparison to larger units, such as words, these subword units also have the advantage of requiring less training data to fully train a speaker

model, thus alleviating the data sparsity problem.

In this work, we investigate subword units that may be suitable to increase speaker recognition performance on heterogeneous noisy data, using audio material from German parliamentary TV broadcasts.

2. Subword Units for Speaker Recognition

A low phone error rate with respect to the actual spoken phones is crucial when using pronunciation features in speaker recognition. In order to reduce the phone error rate on heterogeneous data, we employ larger decoding units and a higher-order language model. We consider *syllables* and *phone clusters derived from syllables* as promising for more robust phonetic speaker recognition on heterogeneous data.

We evaluated syllables for different speech-related tasks such as spoken term detection [10] and pronunciation modelling [11] on German broadcast data. Compared to phone recognition, syllable decoding is inherently more constrained since the pronunciation of each syllable typically covers multiple phonemes. On the evaluation set of [9] (which contains studio quality planned and spontaneous speech from the broadcast domain), our recogniser with a 2-gram *phone* language model achieves a phoneme accuracy of 70%, while a 4-gram *syllable* language model gives a phoneme accuracy of 83%. Although the accuracy is measured for the "ideal pronunciation", i.e. on a phonemisation of the word transcript, and not the actual pronunciation, we assumed that pronunciation-based speaker recognition would benefit from this increased robustness. Syllables still encode variations in pronunciation and should thus be a suitable subword unit for speaker recognition.

As there are considerably more syllables than phones (10k syllables in our syllable lexicon for German vs. 48 phones that the phone recogniser uses), data sparsity becomes more problematic during speaker model training. Therefore, we did not use the recognised syllables directly but subsegmented them into phones and phone clusters, which capture a level of granularity between syllables and phones – 237 phone cluster types occur in our dataset. This enables us to use the benefits of stronger constraints imposed on syllable recognition and at the same time keep the number of subword types small.

2.1. Sub-syllabic Phone Clusters

In [10] we proposed to predict syllable pronunciation variation by exploiting sub-syllabic phone clusters, so-called *position-specific clusters* (PSC). Syllables naturally consist of three phone cluster positions, onset, nucleus, and coda. Greenberg [12] showed that pronunciation variation is predominantly realised at this cluster level. For example, the process of omitting a phone in a word or syllable, known as deletion, tends to occur at the end of the word or syllable. Consider the German function word /U_n_t/ (and). This syllable contains an empty Onset cluster, a vowel in the Nucleus cluster and the two consonants /n_t/ in its Coda cluster. The Coda cluster is, especially in spontaneous speech, often reduced to /n/, thus deleting the syllable-final plosive. By breaking down recognised syllables into their phone clusters one can effectively model regular pronunciation variations observed within syllables. At the same time, the training data can be used more efficiently, as one does not need to observe the pronunciation variants in all syllable contexts, but rather learns variation patterns from sub-syllabic phone clusters. Coming back to the example, we are able to observe phonological processes inherent to pronunciation mannerisms for a given speaker by using phone clusters as identification features.

To break a syllable down into PSCs, a deterministic parser needs to find the cluster boundaries in the syllable. Admissible phone clusters can be segmented following the sonority principle, which states that different phone classes have different sonority values. Sonority reflects the resonance of a phone, for example plosives have the lowest sonority value and vowels the highest. In German syllables, sonority rises from the boundaries of the syllable towards the nucleus, where it reaches its peak. The parser assigns a predefined sonority value to each phone found in the syllable. In the example mentioned above, the vowel /U/ forms the sonority peak with a value of 4, whereas the nasal /n/ and the plosive /t/ are assigned values of 2 and 1, respectively (we use four levels of sonority, with 4 being the highest and 1 the lowest). Once the sonority peak has been identified, for example at a vowel, diphthong, or syllabic consonant, the phones to its left will be categorised as the onset cluster while everything to its right will be the coda cluster. Ulti-

Syllable	Syll w/ sonority values	PSC analysis
S t r U m p f (*sock*)	S[1] t[1] r[3] U[4] m[2] p[1] f[1]	S t r[ONS] U[NUC] m p f[COD]
b r aU x s t (*need*)	b[1] r[3] aU[4] x[1] s[1] t[1]	b r[ONS] aU[NUC] x s t[COD]
S t i: 6: (*bull*)	S[1] t[1] i:[4] 6:[4]	S t[ONS] i: 6:[NUC] ∅[COD]

Table 1: Example of PSC segmentation with sonority.

mately, each PSC contains the clustered phones as well as its position in the syllable. To illustrate the process, consider the examples depicted in Table 1. The second column shows each phone along with its assigned sonority value. Because the vowels and the diphthong both represent the peaks in the syllables, they form the nuclei. Correspondingly, the remaining clusters form the onsets and codas. If more than one phone (such as two consecutive vowels) has the highest value, they are clustered together.

For our experiments, we used the phone clusters with and without the position information appended to each cluster token (yielding tokens like `S_t_r_[ONS]` vs. just `S_t_r`). Using position information may capture speakers' pronunciation mannerisms better – like omitting a *d* in the coda. On the other hand, it again increases the problem of data sparsity. See section 5 for the results.

3. Speaker Modelling

The speaker recognition modelling method used is similar to [3] and [13] and is described in more detail in [14]. Relative frequencies of subword n-grams in the speaker and background training material are computed and combined into log-likelihood ratios $\lambda_i(n)$, which indicate the speaker information for speaker i contained in n-gram n:

$$\lambda_i(n) = log\left[\frac{H_i(n)}{N_i}\right] - log\left[\frac{H_{BG}(n)}{N_{BG}}\right], \quad (1)$$

where $H_i(n)$ and $H_{BG}(n)$ are the number of occurrences of n-gram n, while N_i and N_{BG} are the total n-gram counts for speaker S_i and the background data. The log-likelihood ratios for all n-grams in the training data form the speaker models.

In order to cope with data sparsity, MAP-adaptation of the speaker models from the background model, similar to [13], is employed. The adapted n-gram counts for speaker i, $H_i(n)$, are derived from the counts in the background model and the actual counts in the speaker's training data $H_i(n)'$:

$$H_i(n) = \alpha \cdot H_i(n)' + (1-\alpha) \cdot H_{BG}(n). \quad (2)$$

During recognition, a speaker's score s_i is calculated by summing the speaker n-gram scores $\lambda_i(n)$ for all n-grams in the test data, weighted according to their number of occurrences $c(n)$:

$$s_i = \frac{\sum_n c(n)\lambda_i(n)}{\sum_n c(n)}. \quad (3)$$

4. Experimental Methodology

The following section describes the evaluation scenario and the ASR system which has been used to generate the speaker pronunciation features.

4.1. Data

The data used for the experiments was selected to contain German spontaneous speech in challenging acoustic conditions, such as background noise, reverberation, and varying microphone settings. It was taken from the German parliament's "Web-TV" service[1], which offers video recordings of German parliamentary speeches searchable by speaker. Fourteen federal ministers (6 female and 8 male) were chosen to be the 7 test speakers and 7 impostors. Audio from 5 video recordings per test speaker, each between 2 and 25 minutes in length, was used as speaker training material. The background models were trained with 300 recordings from 300 background speakers. The development and test material consisted of another 2 and 5 videos per test speaker and per impostor, respectively, each from 9 minutes to 1 hour in duration.

For phone and syllable recognition, the audio material was automatically segmented and speech

[1] http://webtv.bundestag.de/iptv/player/macros/bttv/index.html

parts were extracted before recognition. As we have no phone / syllable / word reference transcription for the data, we could not measure the phone / syllable / word error rates of the recognisers. For the spectral system trained as a comparison, at least 6 minutes of audio material from 5 recordings per speaker were manually selected to ensure correct segmentation and speech extraction. Also, speech samples from 64 women and 41 men (of the 300) were manually annotated, yielding at least 1 hour of background training material per gender.

4.2. ASR setup

We used triphone acoustic models optimised for German broadcast news as described in [9]. Due to the acoustic mismatch between the models and the evaluation data, we expect a positive influence of a more constrained language model on the ASR error rate.

For the rather weakly constrained phone recognition used as the baseline, we trained a bigram phone language model using a large corpus of German newswire data which consists of about 150 million running words. The more strongly constrained syllable recogniser used a 4-gram syllable language model built on the same corpus of German newswire data with a 10k syllable dictionary. We applied Good-Turing discounting to smooth the resulting model. The syllable transcripts produced by the 4-gram syllable decoder were broken down by a parser into two types of pronunciation features: strongly constrained phones and phone clusters. The former were used in order to directly compare phone features from weakly constrained phone decoding to phone features from more strongly constrained syllable decoding.

4.3. Speaker Recognition Setup

We trained several speaker recognition systems based on phone n-grams from weakly constrained phone recognition, with modelling as described in Section 3. The MAP-adaptation coefficient α for Equation 2 was empirically determined on the development data to be 0.98. We decided to pick the best of those systems as baseline for further comparisons with the new subword features. See the next section for the results of the baseline systems.

Additionally, we used a basic spectral Gaussian Mixture Model (GMM) system with 512-mixtures per speaker model and a Universal Background Model (UBM) (similar to [1]) as a comparison.

For the following experiments, all speaker and impostor test files were scored against all true speaker models, and the resulting scores were used to produce equal error rates (EERs) and detection-error trade-off (DET) curves to measure the recognition performance.

5. Results

In this section, we present the results of using different subword-based n-gram features for speaker recognition: weakly constrained phones, strongly constrained phones obtained from syllable decoding, and strongly constrained phone clusters.

Preliminary experiments were done to determine which n gives the best performance for our setup with weakly constrained phone n-grams, in order to set the baseline. See Table 2 for the equal error rates. 3-grams performed best, which is consistent with the findings in [13], so we decided to use 3-grams of all subword features in the comparisons.

System	EER
Phone 1-grams	25.7%
Phone 2-grams	20.0%
Phone 3-grams	*14.3%*
Phone 4-grams	20.0%
Phone 5-grams	26.0%

Table 2: Equal error rates for the phone n-gram baseline systems using a 2-gram phone language model for phone recognition.

We then set up speaker recognition systems with n-grams of the different subword features. Figure 1 shows the DET curves for 3-grams of the compared subword units, and Table 3 lists the equal error rates for the tested systems. The baseline of weakly constrained phones was improved upon by using syllable recognition instead of phone recognition: We performed syllable recognition with a syllable 4-gram LM and broke the syllables down into phones. The strongly constrained phone features are more stable with respect to decoding errors on heterogeneous data, improving the EER to 11.8%. Converting the syllables into phone clusters instead of

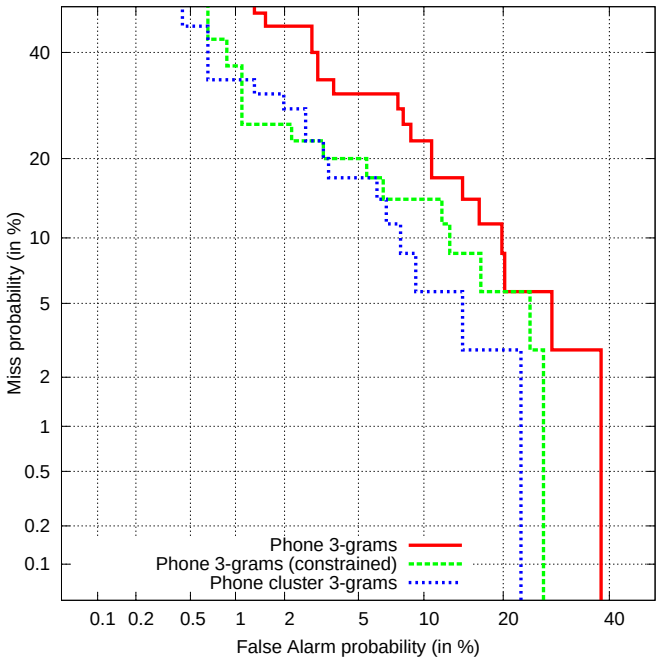

Figure 1: DET-curves for phone 3-grams, phone 3-grams from constrained recognition (with 4-gram syllable LM), and phone cluster 3-grams.

phones further decreases the EER to 8.6%, even below the basic GMM-UBM baseline. The gain may be a result of the larger context captured by phone clusters: as the average cluster contains more than one phone (an average of 1.22 phones per cluster for the data used), we effectively model variable length phone n-grams. Phone cluster 3-grams can span more than 3 phones, and thus correspond to full syllables in many cases (the average number of phone clusters per syllable was 2.25). Strongly constrained phone 4-grams alone, however, result in a worse performance, with an ERR of 11.4% (constrained phone 5-grams have an EER of 17.3%). A reason why clusters outperform single phones could be their variability in terms of context length. Instead of using simple n-grams of phones we rely on the linguistically determined cluster lengths inherent to syllables, which means that cluster trigrams model more context than phone trigrams.

Finally, we tested phone clusters with and without position information and found that position information does not improve speaker recognition performance on the given data, which, again, may be due to data sparsity.

System	EER
Spectral (GMM-UBM)	11.4%
Phone 3-grams baseline (2-gram phone LM)	14.3%
Phone 3-grams (4-gram syllable LM)	11.8%
Phone cluster 3-grams (4-gram syllable LM)	*8.6%*
Phone 4-grams (4-gram syllable LM)	11.4%
Phone 5-grams (4-gram syllable LM)	17.3%
Phone cluster with position information 3-grams (4-gram syllable LM)	8.8%

Table 3: Equal error rates for the evaluated systems.

6. Conclusion

We investigated the use of different subword units and a more strongly constrained ASR decoding process for robust pronunciation-based speaker recognition on challenging spontaneous speech data. We found that a strong language model indeed improves speaker recognition performance on real-life data, lowering the decoding error caused by data heterogeneity while at the same time capturing speaker specific pronunciation variation. Also, using subsyllabic phone clusters instead of phones as features further enhances results, decreasing the EER from 14.3% for the baseline phone features to 8.6% for phone clusters with a strongly constrained decoding process.

In the future, we plan to evaluate speaker recognition with phone clusters with different modelling methods on larger datasets and in languages other than German. Moreover we aim at fusing the proposed method with other approaches.

7. Acknowledgements

This study was funded by the CONTENTUS scenario of the THESEUS project, the EU IST VITALAS project, and the SMUDI project.

8. References

[1] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.

[2] D.A. Reynolds, J.P. Campbell, W.M. Campbell, R.B. Dunn, T.P. Gleason, D.A. Jones, T.F. Quatieri, C.B. Quillen, D.E. Sturim, and P.A. Torres-Carrasquillo, "Beyond cepstra: Exploiting high-level information in speaker recognition," in *Workshop on Multimodal User Authentication*, Santa Barbara, California, December 2003, pp. 223–229.

[3] W.D. Andrews, M.A. Kohler, J.P. Campbell, J.J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *ICASSP '02*, 2002, vol. 1, pp. 149–152.

[4] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones, and T.R. Leek, "Phonetic speaker recognition with support vector machines," in *Neural Information Processing Systems Conference*, Vancouver, British Columbia, December 2003, pp. 1377–1384.

[5] Q. Jin, J. Navratil, D.A. Reynolds, J.P. Campbell, W.D. Andrews, and J.S. Abramson, "Combining cross-stream and time dimensions in phonetic speaker recognition," in *ICASSP '03*, 2003, vol. 4, pp. 800–803.

[6] D. Klusacek, J. Navratil, D.A. Reynolds, and J.P. Campbell, "Conditional pronunciation modeling in speaker detection," in *ICASSP '03*, 2003, vol. 4, pp. 804–807.

[7] T. Bocklet and E. Shriberg, "Speaker recognition using syllable-based constraints for cepstral frame selection," in *ICASSP '09*, 2009, pp. 4525–4528.

[8] E. Shriberg, L. Ferrer, A. Venkataraman, and S. Kajarekar, "SVM modeling of "SNERF-grams" for speaker recognition," in *International Conference on Spoken Language Processing*, Jeju, Korea, 2004, pp. 1409–1412.

[9] T. Mertens and D. Schneider, "Efficient subword lattice retrieval for german spoken term detection," in *ICASSP '09*, 2009, pp. 4885–4888.

[10] T. Mertens, D. Schneider, and J. Köhler, "Merging search spaces for subword spoken term detection," in *Interspeech 09*, 2009, pp. 2127–2131.

[11] T. Mertens, D. Schneider, A. Næss, and T. Svendsen, "Lexicon adaptation for subword speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU '09)*, 2009.

[12] S. Greenberg, "Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[13] B. Baker, R. Vogt, M. Mason, and S. Sridharan, "Improved phonetic and lexical speaker recognition through MAP adaptation," in *ODYSSEY04 - The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004, pp. 91–96.

[14] D. Baum, "Topic-based speaker recognition for German parliamentary speeches," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU '09)*, 2009.