Phonemic Suitability in Text-Dependent Speaker Verification

Bianca Aschenberner, Christian S. Pilz

VOICE.TRUST AG,
Geisenhausener Str.15, D-81379 München, Germany
bas@voicetrust.com, cpi@voicetrust.com

Abstract

In this paper, phoneme suitability within text-dependent speaker verification is investigated. It is assumed that using certain phonemes and avoiding others improves the recognition accuracy. According to a supposed definition of phonemic suitability ratio, the phoneme selection was applied to the utterances of the Voice Trust CC speech database. It could be shown that an efficient selection of password concepts, which is, avoiding not-suitable and preferring suitable phonemes, yields to better recognition results in a text-dependent speaker verification. Besides the improved recognition accuracy, this technique is very useful to speed up the whole authentication process, thus raising the system's usability and user-friendliness in customer place, by avoiding the insufficient speech sounds in advance. Furthermore, the phoneme suitability and according utterance selection increases the security by a priori removing those utterances which do not gain an adequate confidence level.
Index terms: speaker recognition, speaker verification, phoneme suitability, phoneme weighting.

1. Introduction

In the last years, extensive research and development has been done in the field of biometric authentication and has already yielded to marketable products, such as iris scan, finger or voice print verification. Finding and using the most sufficient and person-unique parameters for such authentications is reasonably important and desirable, not only for best verification accuracy, but also for highest user acceptability by offering concise, time-saving processes of extracting and computing the person-specific parameters.
Regarding particularly voice biometric applications and their performance, the corresponding aim is to find speech samples with a maximum of information content and minimum of time consumption for enrolling and verifying the user.
However, in practical examination of the authentication accuracy when random speech samples are used, the following trade-off becomes obvious: either the security level is increased, then at the expense of the required duration, or the time amount is decreased, then at the expense of security.
But it could be proven, that both aims can be converged by using certain pre-selected speech data. What speech content this is, how it should be included in the enrolment and verification process, and to what amount this improves the authentication performance shall be explained in this paper.

2. Examination of Voice Characteristics

The following section gives some brief introduction to the analysis of voice characteristics and their classification into phonemes. Further, it displays the correlation of speech sounds and their respective recognition accuracy.

2.1. General description

The basis of the analyzing speech sounds, extracting their features and further processing them are the physical characteristics of the phonemes. As it is well known, these can be gained by modelizing the speech production from glottis to mouth radiation as a one-end-closed, one-end-opened tube with different cross-sections. At the glottis, either the vocal folds stay open and do not alter the pulmonal air stream, or phonation is produced by the vocal folds' vibration. If phonation is missing, the articulators' location and manner of air constriction produces the according sounds, such as unvoiced fricatives or plosives. If phonation is given, the vocal tract's shape alters the wave propagation so that the according voiced sounds are produced, such as vowels.
All together, the whole phoneme inventory can be described by modeling the glottis, the vocal tract and the articulators, and by thus gaining the according physical equivalents, such as their frequency and amplitude.

2.2. Voice Parameters for Speaker Verification

The foundation of an elaborated verification engine is a well-generated speaker model and a proper verification against these previously gained and processed speaker data. Therefore, the engine utilizes the specific physical characteristics of the speech sounds produced by the particular speaker and generates the enrolment model.
But not all of the extractable sound features are same usable for a well speaker verification, as also [6] found when analyzing different sonagrams. As a first demur, unvoiced sounds are only random noises with a certain frequency spectrum and amplitude, they do not carry much information about the speaker. In contrast, voiced sounds allow the inference to the speaker's vocal tract, his vocal folds and articulators' behavior by analyzing the according wave propagation and shaping. But again, not all voiced phonemes are necessarily same usable for a speaker authentication, as several researchers testified. The performance of speaker authentication systems is displayed by their EER (equal error rate), representing the intersection of false acceptances and false rejections. [1] found the following EERs when analyzing the speech recognition of a certain set of phonemes:

Phoneme SAMPA	EER	Phoneme SAMPA	EER
a:	8,2%	k	23,7%
m	8,5%	v	24,7%
N	9,7%	t	25,3%
E	10,6%	n	31,0%
f	21,0%		

Table 1. *EERs of certain phonemes during speech recognition*

Even though the whole phoneme inventory was not covered, [4] calculated, among others, following further ERRs:

Phoneme SAMPA	**EER**	**Phoneme SAMPA**	**EER**
i:	10,5%	s	18,6%
o:	11,4%	S	15,7%
u:	18,1%		

Table 2. *EERs of further phonemes in speech recognition*

[7] proved in another extensive examination of these findings and their use for speaker verification, that the calculated EERs hold not only for speech, but also for speaker recognition.
When comparing these results to the findings of other researchers though, such as [2] or [3], some ambiguities about the suitability of certain phonemes occur. For example, researchers disagree about the suitability of the nasal /n/. Besides, no complete evaluation of the whole phoneme inventory and its suitability for speaker authentication could be found. So it would be advisable to prepare another extensive analysis about all existing speech sounds and their contribution to speaker authentication. But since the quoted work [4] seems to be the most elaborated and the most complete study about this issue, these findings were used for this paper's further research and analysis.
And generally, research agrees: the different speech sounds differ in their suitability for speaker authentication. Thus, they should be treated differently before or during a speaker verification, in order to increase the verifier's EER, to increase the security level and to decrease the time consumption of the enrolment and verification procedure.

3. Phoneme (Pre-)Selection for Speaker Authentication

As described, different speech sounds have a different impact on the performance of speech recognition systems, being displayed by their EERs. How that impact can and should be differentiated and how different phoneme suitabilities can be derived shall be explained in the following.

3.1. General Notes

Speaker authentication systems need to account for security and best performance, but also for user-friendliness. At this, the users cannot be asked to separately pronounce best suitable phonemes to enroll or verify. Instead, the user will want to speak whole phrases, prompted by the text-dependent speaker verification system.
For the verification procedure, the phoneme classes should be favored in following ranking for best verification accuracy:

1) Vowels
2) Nasals, diphthongs and glides
3) Fricatives
4) Plosives

At this, it is advisable to prefer those phrases for both enrolment and verification which contain a maximum of vowels and a minimum of plosives.
This could be achieved by either pre-selecting the prompts which the user shall repeat, or by segmenting the given speech input into its phonemes and to only pass those ones through which are well suitable. The latter option seems to be more critical and laborious, regarding segmentation difficulties, coarticulation effects, and especially regarding the fact that sometimes, a priori knowledge about the phonetic content is not available. At this, it would be more advantageous to pre-select the phonemes and prompts so that the general accuracy is increased and the required process time is decreased.

3.2. Aspects of Phoneme Weighting

Regarding the phonemic suitability, two basic ways of evaluating the phoneme inventory and its usability are suggested: (1) the general classification into suitable and not-suitable phonemes, (2) some further weighting of the sounds according to their empirical EERs. The following section discusses the two considerations.
The first evaluation just defines a certain EER above which the phonemes are well suitable and below which the phonemes should not be taken into account. So the suitable phonemes are weighted with the factor 1 and are thus given highest importance for speaker verification. The others are weighted with the factor 0 and are thus not taken into the authentication decision. The overall weighting of the according word is then calculated as follows:

$$suitability\ ratio = \frac{M}{N} \qquad (1)$$

Here, N represents the total number of phonemes and M displays the number of suitable phonemes.
As an example, following phoneme substitutes and their fictitious EERs are assumed, and it is suggested to only take those phonemes into account which have an EER above 0,7, as presented in the following table 3.

Phoneme	Fictitious EERs	Suitability
A	0,9	Suitable
B	0,6	-
C	0,4	-
D	0,5	-
E	0,8	Suitable

Table 3. *Phoneme substitutes, their fictitious EER and their according suitability*

Using this suitability differentiation, the symbol sequence "CEABDE" would result in the ratio 0,5, according to the above given formula for the weighted suitability. Another sequence "CEABDA" would also have a phonetic suitability of 0.5.
Now, it shall be assume that the empirical EERs are taken into account, according to table 3, not only the differentiation of suitable (weight 1) and unsuitable (weight 0). The EERs-including suitability ratio will then be calculated as follows:

$$weighted\ suitability\ ratio = \frac{1}{N}\sum_{i}^{N} a_i \qquad (2)$$

N represents the total number of phonemes and a_i displays the weighting (such as 0,8 for D), according to the empirical recognition accuracy obtained by the EER.
Calculating the suitability now by using this second, weighted suitability ratio, the above mentioned symbol sequences will differ in their suitability: The first sequence "CEABDE" would have a suitability of 0,667. In contrast, the second sequence "CEABDA" – which has the same suitability 0,5 when calculating the ratio by the first formula – will be more suitable when determined by formula two, now having a ratio of 0,683.

Obviously, these two different approaches of defining the suitability ratio for speaker authentication lead to different results of phoneme suitabilities.
The latter one seems to be less practicable due to incompleteness of empirical phoneme EERs and their dependency on the database in use, so the generalization ability seems to suffer.
Hence, the unweighted proposal should be preferred and shall be used and analyzed exclusively in the following experiments.

4. Experimental Setup

In this section, we experimentally investigate the efficiency of the phonetic suitability on a text-dependent speaker verification task.

4.1. Used Database

The phonetic suitability technique is evaluated on the 350 speaker cellular telephone speech database, developed and used by the Voice Trust AG for the Common Criteria certification from 2005. Each speaker is asked to repeat five different German name pairs, a generic user ID, a generic name pair and a generic pass phrase for six times each. The first four repetitions are used to train speaker models, the other repetitions are used for verifications.
The following table shows the used speech data, whereas the corresponding phonemic transcriptions are determined manually:

	Orthogr.	**Phonemic**
Cr1	Rosemarie Maximilian	ro:z@mari: maksImi:lIa:n
Cr2	Lieselotte Sebastian	li:z@lOt@ z@bastIa:n
Cr3	Veronika Ferdinand	ve:ro:nIka fErdi:nant
Cr4	Evamaria Konstantin	e:famari:a kOnstanti:n
Cr5	Christiane Dagobert	krIstIa:n@ da:go:bErt
GID	GHI456	ge:ha:i:fi:rfYnfsEks
GN.	Karoline Mustermann	karo:li:n@ mUst@rman
GPhr.	Meine Stimme ist mein Passwort.	maIn@StIm@Ist maInpasvOrt

Table 4. *Database's utterance concepts and corresponding SAMPA transcriptions*

The experimental database consists of a gender-balanced subset of 100 speakers of the Voice Trust Common Criteria corpus. Also, four samples of all concepts were used as training repetitions, and one repetition was used as the verification basis.

4.2. Experimental Conditions

The text-dependent speaker verification system which is used for this paper's experiments is based on a whole word DTW-HMM (dynamic time warping – hidden markov model) classifier. The feature vector consists of twelve PFLPCCs [5] and twelve □□□□□□□□ (pole-filtered linear prediction cepstrum coefficients) being extracted from energy based silence removed frames of 30-ms length with 20-ms frame shift.
Acoustic models are trained on DTW warped cepstra, whereby the utterance with the smallest length was chosen as the reference template for the according warping. The feature vectors of training data are pooled to HMM states, using Viterbi alignment. Each HMM state has two Gaussian mixture components with diagonal covariance matrices.
During verification, the decision is made on the normalized average loglikelihood. The likelihoods were normalized by a sigmoid function and scaled to a range from –1 to 1, in order to get a zero boundary trade-off when determining impostor and speaker mean on a fixed 30 speaker cohort set during training.

4.3. Experimental Assumption

The suitable phonemes where defined by referring to the findings of [4] and [7]. Only the phonemes with an EER below about 12 % were taken into account, plus the fricative /S/ which was assumed to be the best sound to be used as a separating phoneme between vowels. Although all vowels are known to carry the most speaker-specific information, vowels such as /u:/ or /9/ were excluded because their EER was either too low, or no data were available. Furthermore, other fricatives but /S/ – such as /x/ whose EER is also rather low – were also excluded from the suitability corpus because of the verification system's premises: the verify application which the evaluations were executed with uses incoming phone calls as the source of speech input. Since the telephone band is likely to be limited to 300-3400 Hz, sounds with their amplitude maximum below or above that frame were avoided.
All in all, following phonemes were defined to be best suitable for improving a speaker verification:
a,a:/ /e,e:/, /E,E:/, /I,i:/, /O,o:/, /Y,y:/, /m/, /N/, /j/, /S/.

4.4. Experimental Results

The evaluation of the speaker verification system is based on the equal error rate (EER) curves which show the tradeoff between the false acceptance (FA) and the false rejection (FR). The EERs were determined on the 100 user crossmatches, and are listed in the table below in descending order. The unweighted phonetic suitability coefficients where also determined according to formula (1) and the chosen suitable phoneme inventory, and are assigned to the corresponding concept.
The results are given as follows:

	EER in %:	Suitability ratio:
CR4:	1,07	0,6667
CR1 :	1,66	0,6316
CR3:	1,95	0,5882
GPhr.:	2,00	0,541
CR2:	2,23	0,3529
GN.:	2,46	0,4706
CR5:	2,55	0,4118
GID:	4,83	0,4375

Table 5. *Determined EERs and suitability ratios.*

In almost all cases, small EERs correspond to high phoneme suitabilities and high EERs correspond to small phoneme suitability. And when comparing the mean EERs of the

concepts with suitabilities above 0,6 and below 0,6, the verification accuracy is improved by about 49%.

5. Conclusions

In this paper, a phonetic suitability is investigated on a text-dependent speaker verification task. It could be proven, that utterance selection based on a phonemic suitability ratio before the actual recognition procedure is strongly enhancing the speaker recognition results. At this, the selected phonemes which were assumed to increase the recognition accuracy are indeed improving the performance when they are preferred as speaker input for the authentication process.
It further research it could still be analyzed though, how the whole phoneme inventory can be classified into suitable and not-suitable sounds, and to what extent the authentication accuracy might be further optimized by this phonemic classification.

6. Acknowledgements

Thanks to the Voice.Trust AG for providing the Common Criteria database for evaluation and to the University of Applied Sciences Friedberg-Giessen for their support.

7. References

[1] Euler, S., Langlitz, R. and Zinke, J., “Compariso of whole word and subword modeling techniques for speaker verification with limited training data”, ICASSP-97, Munich, 1997.

[2] Fattah, M. A., Ren, F. and Kuroiwa, S., “Phoneme based Speaker Modeling to Improve Speaker Identification”, AIML 06 International Conference, 13-15 June 06, Sharm El Sheikh, Egypt, 2006.

[3] Heuvel, H. van den, “Speaker Variability in Acoustic Poroperties of Dutch Phoneme Realisations”, Dissertation, University of Nijmegen. Netherlands, 1996.

[4] Langlitz, R., “Sprecherverifikation mit Hidden Markov Modellen”, Diploma Thesis, University of Applied Sciences Giessen-Friedberg, Friedberg, 1996.

[5] Naik., D., “Pole-filtered cepstral mean subtraction”, Acoustics, Speech, and Signal Processing, ICASSP-95, vol.1, 157 – 160, 1995.

[6] Savic, M., Gupta, S. K., “Variable parameter speaker verification system based on hidden Markov modeling.” Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 281-284, 1990.

[7] Varvelli, G., “Aspekte zur Sicherheit in sprachbasierten biometrischen Sicherheitssystemen”, Diploma Thesis, University of Applied Sciences Giessen-Friedberg, Friedberg, 2002.