Viseme Recognition System Based on Transformed Acoustic Models

A. Zgank[1], Z. Kacic[1]
[1]*University of Maribor, Institute of Electronics and Telecommunications,*
Smetanova St. 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si

Abstract**—Viseme recognition from speech is one of the methods needed to operate a talking head system, which can be used in various areas, such as mobile services and applications, gaming, the entertainment industry, and so on. This paper proposes a novel method for generating acoustic models for viseme recognition from speech. The viseme acoustic models were generated using transformations from trained phoneme acoustic models. The proposed transformation method is language-independent; only the available speech resources are needed. The viseme sequence with corresponding time information was produced as a result of recognition using context-dependent acoustic models. The evaluation of the proposed acoustic models' transformation method was carried out on a test scenario with phonetically balanced words, in which the results were compared to the baseline viseme recognition system. The improvement in viseme accuracy was statistically significant when using the proposed method for transforming acoustic models.**

Index Terms**—Automatic speech recognition, hidden Markov models, human computer interaction, viseme modeling.**

I. Introduction

Advanced human computer interfaces may include a virtual assistant [1] to achieve natural communication with the user. The virtual assistant is frequently visualized by facial animation in the form of a talking head [2], [3]. The perceptual quality of human-computer interaction can be improved if gestures [4] and various other characteristics of communication [5], [6] are included in the virtual assistant.

Facial animation applies visemes as a basic speech unit. A sequence of visemes with appropriate time boundaries is needed to model the movement of a talking head's mouth [7], [8]. There are two main approaches for creating this audio-to-visual conversion. If a speech synthesis module is used for generating the speech signal, a viseme sequence with time boundaries can be produced using mapping from the phoneme sequence, which is usually an intermediate result of a speech synthesis algorithm [9]. If the recorded or live speech signal is used for the talking head's spoken modality, viseme recognition must be carried out in order to produce a viseme sequence with time information [10]. The research presented in this paper focuses on the second case, with emphasis on the acoustic modeling of visemes in speech. The baseline approach for viseme recognition from speech uses acoustic models, in which the models are trained as visemes from the initialization onwards. The main disadvantage of this approach is that one viseme usually covers more than one phoneme. As a result, it is therefore difficult to obtain a satisfactory model of acoustic-phonetic space.

This paper proposes a new viseme acoustic modeling method, in which visemes are transformed from phoneme acoustic models and trained through subsequent steps as context-dependent viseme acoustic models. The proposed acoustic models' transformation method is language-independent, and can be used for any language with available speech resources. The experiments for evaluating the proposed method were carried out for Slovenian using a 1000 FDB SpeechDat(II) database.

The paper is organized as follows. The theoretical background on the transformation from phoneme to viseme acoustic models is given in Section II. The speech database is described in Section III. An overview of the experimental setup is presented in Section IV. The results are given in Section V, and the conclusion in Section VI.

II. Viseme Speech Recognition

A viseme is a basic unit during visual speech processing [11]. It groups together within a speech signal all those phonemes that share the same mouth/lip shape (Fig. 1). Visemes as basic units of acoustic models represent much larger acoustic-phonetic space within the speech signal than phonemes, which are typically used as basic units in acoustic models for speech recognition. The main idea of the proposed method is to generate viseme acoustic models through transformation of phoneme acoustic models. It can be expected that the transformed acoustic models will be better suited for the specific recognition task. The transformation into viseme acoustic models is defined as a combination of adequate phoneme acoustic models.

A single Gaussian continuous density HMM acoustic model has three different types of parameters: mean values, variances, and transition probabilities. The transformed viseme acoustic model V is defined as

$$V = \sum_{i=1}^{I_P} w_i P_i \ , \qquad (1)$$

Manuscript received February 28, 2013; accepted June 4, 2013.
This work was partially supported by Slovenian Research Agency under contract number P2-0069 "Advanced methods of interaction in telecommunication".

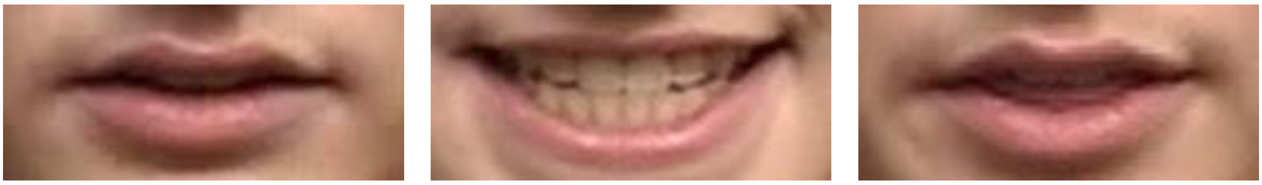

Fig. 1. Examples of visemes from the CUAVE multimodal (audio/video) speech corpus [12].

where P_i denotes a particular phoneme acoustic model and w_i denotes the weight calculated for this phoneme acoustic model. The variable i denotes the current phoneme acoustic model from the pool of individual visemes. I_p is limited to the number of all phonemes that share the same viseme.

The means of the transformed viseme acoustic model V are defined as

$$\mu_V = \sum_{i=1}^{I_p} w_i \mu_{P_i}, \tag{2}$$

where μ_V and μ_{Pi} denote the mean values of the transformed viseme and source phoneme acoustic models, respectively. The weight is represented by w_i.

The transformed viseme variances are set to the maximal variance for all the involved phoneme probability density functions

$$\nu_V = \max_i(\nu_{P_i}), \tag{3}$$

where ν_V denotes the variances of the transformed viseme acoustic model and ν_{P_i} denotes the variances of the phoneme acoustic models, respectively.

The last element of the transformed viseme acoustic model is a matrix with transition probabilities. The transformed transition probabilities are defined as

$$\alpha_V = \sum_{i=1}^{I_P} w_i \alpha_{P_i}, \tag{4}$$

where α_V denotes the transformed viseme transition probability and α_{P_i} denotes the transition probabilities of the phoneme acoustic models, respectively. The variable w_i represents the weight.

The influential weight w_i in (1), (2), and (4) can be estimated using various approaches and metrics. In our case, the weight w_i is defined as the proportion of the available spoken training material for each phoneme acoustic model, which is defined as

$$w_i = \frac{n_i}{\sum_{j=1}^{I_P} n_j}, \tag{5}$$

where n_i denotes the number of feature frames per particular phoneme acoustic model and the sum of n_j denotes the number of feature frames for all those phoneme acoustic models that belong to the same viseme.

The advantage of the proposed transformation method is that existing phoneme acoustic models can be reused as a starting point for building viseme acoustic models. In this way, those available acoustic models potentially trained with complex algorithms can be re-used. The complexity of transformed viseme acoustic models is similar to those of baseline viseme acoustic models and smaller than if the phoneme acoustic models combined with phoneme-to-viseme mapping during a post-processing step were to be used for viseme recognition. The lesser complexity of acoustic models results in faster recognition and less memory resources needed for operating the system. The level of available system resources is especially important in the case of a system with complex processing and modeling tasks, as is frequently the case when using talking heads and live speech recognition [4].

III. SPEECH DATABASE

The experiments were carried out using the Slovenian 1000 FDB SpeechDat(II) database, which is part of the SpeechDat family with more than 50 different languages available. A database of this type can be used to develop voice-driven services [13] and various applications. It was recorded over a fixed telephone with 1,000 different speakers within the set. The training set had 800 speakers; the remaining 200 were used for evaluation. During the designing of the database, special care was taken regarding the speakers' demographic characteristics to ensure that the set was in optimum balance. Each recording session covered 43 different utterances per speaker.

The number of basic units within the speech database was important from the point of view of acoustic modeling [14]. There were 46 different phonemes in the Slovenian SpeechDat(II) database. Only the 27 most frequent Slovenian phonemes were used during the experiments. The remaining 19 phonemes were mapped into the more similar ones using phonetic expertise. The original speech database transcriptions were based on the phonemes. The appropriate viseme transcriptions were created by applying phoneme-to-viseme mapping during the training pre-processing step.

The MPEG-4 standard [11] defines 16 different visemes. The list of Slovenian phonemes and their mapping into MPEG-4 visemes is shown in Table I. The visemes that included the most phonemes were “PP”, “FF”, “kk”, “CH”, “SS”, and “nn”. The more specific Slovenian visemes – for example, “J” – included only one phoneme. The viseme “silence” was used to model long and short pauses in the speech signal.

The evaluation of a speech recognition system can be done using various isolated and connected word scenarios: digits, numbers, persons' names, city names, command words, and so on. The SpeechDat(II) subset of phonetically balanced isolated words (W set) was used for this purpose. Each speaker in the database uttered four different words.

There were a total of 1,491 different words within the set. From the acoustic modeling point of view, such a test set is very suitable for evaluation because the phonemes of the language are correctly represented within it.

TABLE I. LIST OF MPEG-4 VISEMES.

MPEG-4 Viseme	**Phoneme**
silence	long pause, short pause
aa	a
E	e
ih	i
oh	o
ou	u
eh	@
RR	@r
PP	p, m, b
FF	f, v
kk	k, g, x
DD	t, d
CH	tS, S, Z
SS	ts, s, z
nn	n, l, r
J	j

The proposed method for viseme recognition was based on acoustic modeling. Therefore an isolated word test set was selected for evaluation and a viseme loop was applied as recognition grammar. Such an approach excluded the influence of the statistical language model on the viseme recognition results.

IV. EXPERIMENTAL SYSTEM

Two different experimental systems were developed. The modified monolingual COST 278 MASPER scripts [15] were taken as a starting point. The first experimental system was used as a baseline. It was completely designed based on viseme acoustic models. The visemes were produced during the first step because the phoneme-to-viseme mapping had already been carried out during the pre-processing of the training transcriptions. The second experimental system introduced the proposed method for transforming the acoustic models. In this case, first the phoneme acoustic models were trained, which were then transformed into viseme acoustic models in the second part of training.

The speech signal was converted into features using the 12 mel-cepstral coefficients and energy; thereafter, first- and second-order derivatives were calculated. The feature extraction window was 25 ms long with a 10 ms time shift. The final feature vector with 39 elements was constructed after cepstral mean normalization.

Both experimental systems were based on the Hidden Markov Models (HMM) automatic speech recognition approach. Acoustic models with three-state left-right topology were used, in which continuous Gaussian probability density functions were applied to each model's state.

The first part of the training procedure, in which a generation of context-independent acoustic models took part [16], was identical for both systems. In the first step, the context-independent acoustic models were initialized with the global values and then trained with six iterations of embedded Baum–Welch re-estimation. The trained acoustic models were included in the forced-realigning procedure with the goal of improving the transcriptions and detecting outliers within the training corpus. The results of the realigning procedure showed that 0.13 % and 0.11 % of training utterances were classified as outliers and, as such, were excluded from the subsequent training procedures.

In the second stage of the first part, the task was to train improved acoustic models for the second round of the forced realigning procedure. The second stage of the training procedure used the improved transcriptions to train a new set of context-independent acoustic models. This time, the initial values of the probability density function and transition matrix were set individually for each acoustic model. The acoustic models were trained in an iterative manner, in which the number of Gaussian probability density functions per state was increased in stepwise. The resulting context-independent acoustic models had 32 Gaussian mixtures per state. A second forced-realigning procedure was carried out, classifying less than 0.07 % and 0.06 % of utterances as outliers.

The second part of the training procedure was diverse for the baseline and the transformed viseme experimental system. In the case of the baseline viseme system, the context-independent acoustic models were directly used to build the context-dependent version. On the other hand, in the case of the transformed viseme system, first the transformation from phoneme to viseme acoustic models was done using the proposed method. The transformed viseme acoustic models were then used to build the context-dependent versions of the models.

The large number of free acoustic models' parameters was controlled by decision tree based state clustering. The increase of log likelihood after node splitting was used as the threshold to guarantee the same complexity for both versions of the acoustic models. The tied state context-dependent acoustic models were then further trained. During this final training, the number of Gaussian mixtures per state was again increased. Acoustic models with three complexities (low: 4, medium: 8, and high: 16 Gaussian probability density functions per state) were created for both versions.

V. RESULTS

The proposed method for the transformation of acoustic models from phonemes to visemes was evaluated in an indirect way, using the viseme recognition results. The results are presented as viseme accuracy (*VA*), which is defined as

$$VA(\%) = \frac{N - S - D - I}{N} \cdot 100, \tag{6}$$

where *N* denotes the number of all visemes, *S* the number of substituted visemes, *I* the number of inserted visemes, and *D* the number of deleted visemes within the test set. Table II presents the viseme accuracy results, in which the baseline viseme acoustic models were used for evaluation.

The baseline viseme acoustic models, trained from the initialization with visemes, achieved 57.99 % viseme accuracy using the low-complexity acoustic models. When the medium- and high-complexity acoustic models were applied, the viseme accuracy improved to 61.95 % and 64.21 %, respectively. This increase in performance is caused be increased complexity of the acoustic models, which can better model the acoustic-phonetic diversities of visemes. The baseline results achieved are comparable with the viseme recognition results of similar systems [17], [18].

TABLE II. VISEME ACCURACY WITH BASELINE VISEME ACOUSTIC MODELS.

System complexity	**Viseme accuracy (%)**	**Confidence interval (%)**
Low	57.99	±2.86
Medium	61.95	±2.82
High	64.21	±2.78

Table III presents the viseme recognition results, in which the transformed visemes were applied for the recognition task.

TABLE III. VISEME ACCURACY WITH TRANSFORMED VISEME ACOUSTIC MODELS.

System complexity	**Viseme accuracy (%)**	**Confidence interval (%)**
Low	61.09	±2.83
Medium	63.80	±2.79
High	66.51	±2.74

The viseme accuracy improved by 3.10 % absolutely when using low-complexity transformed viseme acoustic models. In the cases of medium- and high-complexity transformed acoustic models, the absolute improvement of viseme accuracy was smaller: 1.85 % and 2.30 %, respectively. The best overall viseme accuracy of 66.51 % was achieved using the high-complexity transformed viseme acoustic models. This improvement reflected the fact that the transformed viseme acoustic models are better suitable for this complex recognition task. A plausible explanation would be that for the transformed acoustic models each viseme was built as a combination of phoneme acoustic models, and could therefore better model the speech signal than when viseme acoustic models were trained from the initialization onwards.

VI. CONCLUSIONS

This paper presented a novel approach for building acoustic models for viseme recognition tasks. The basic idea was to transform the existing phoneme acoustic models into visemes by taking into account the characteristics of the training data. The results of experimental evaluation show the benefits of transformed viseme acoustic models because statistically significant improvements in accuracy were achieved. The highest improvement was 3.10 % absolutely for the low-complexity models, and the best overall viseme accuracy was 66.51 % with the high-complexity models.

Our future work will be oriented towards improving methods for calculating the weights w_i. Another possibility of increasing the performance of a viseme recognition system would be in the area of a decision tree–based clustering algorithm.

REFERENCES

[1] T. Kosutic, M. Mosmondor, I. Andrisek, M. Weber, I. S. Pandzic, M. Matijasevic, “Personalized avatars for mobile entertainment”, *Mobile Information Systems*, vol. 2, no. 2–3, pp. 95–110, 2006.

[2] I. Mazonaviciute, R. Bausys, “English talking head adaptation for Lithuanian speech animation”, *Information Technology and Control*, vol. 38, no. 3, pp. 217–224, 2009.

[3] I. Mlakar, M. Rojc, “Personalized expressive embodied conversational agent EVA”, in *Proc. of Int. Conf. Visualisation, Imaging and Simulation*, Faro, 2010.

[4] G. Zoric, R. Forchheimer, I. S. Pandzic, “On creating multimodal virtual humans – real time speech driven facial gesturing”, *Multimedia Tools and Applications*, vol. 54, no. 1, pp. 165–179, 2011. [Online]. Available: http://dx.doi.org/10.1007/s11042-010-0526-y

[5] D. Verdonik, “Between understanding and misunderstanding”, *Journal of Pragmatics*, vol. 42, no. 5, pp. 1364–1379, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.pragma.2009.09.007

[6] K. Vicsi, G. Szaszak, “Using prosody to improve automatic speech recognition”, *Speech Communication*, vol. 52, no. 5, pp. 413–426, 2010. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2010.01.003

[7] M. Lehr, A. Arruti, A. Ortiz, D. Oyarzun, M. Obach, “Speech driven facial animation using HMMs in Basque”, in *Proc. of TSD 2006, LNAI 4188*, 2006, pp. 415–422.

[8] G. Zoric, I. S. Pandzic, “Real-time language independent lip synchronization method using a genetic algorithm”, *Signal Processing*, vol. 86, no. 12, pp. 3644–3656, 2006. [Online]. Available: http://dx.doi.org/10.1016/j.sigpro.2006.02.038

[9] M. Rojc, Z. Kacic, “A unified approach to grapheme-to-phoneme conversion for the plattos Slovenian text-to-speech system”, *Applied Artificial Intelligence*, vol. 21, no. 6, pp. 563–603, 2007. [Online]. Available: http://dx.doi.org/10.1080/08839510701409086

[10] I. Mazonaviciute, R. Bausys, “Translingual visemes mapping for Lithuanian speech animation”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering)*, no. 5, pp. 95–98, 2011.

[11] I. S. Pandzic, R. Forchheimer, *MPEG-4 Facial Animation – The Standard, Implementations and Applications*. John Wiley & Sons, 2002. [Online]. Available: http://dx.doi.org/10.1002/0470854626

[12] E. K. Patterson, S. Gurbuz, Z. Tufekci, J. N. Gowdy, “Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus”, *EURASIP Journal on Advances in Signal Processing*, no. 11, pp. 1189–1201, 2002. [Online]. Available: http://dx.doi.org/10.1155/S1110865702206101

[13] R. Maskeliunas, K. Ratkevicius, V. Rudzionis, “Voice-based human-machine interaction modeling for automated information services”, *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering)*, no. 4, pp. 109–112, 2011.

[14] N. Theera-Umpon, S. Chansareewittaya, S. Auephanwiriyakul, “Phoneme and tonal accent recognition for Thai speech”, *Expert Systems with Applications*, vol. 38, no. 10, pp. 13254–13259, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2011.04.142

[15] A. Zgank, Z. Kacic, F. Diehl, K. Vicsi, G. Szaszak, J. Juhar, S. Lihan, “The COST 278 MASPER initiative: Crosslingual speech recognition with large telephone databases”, in *Proc. of LREC 2004*, Lisbon, 2004.

[16] B. Dropuljic, D. Petrinovic, “Development of acoustic model for Croatian language using HTK”, *Automatika*, vol. 51, no. 1, pp. 79–88, 2010.

[17] P. Lucey, T. Martin, S. Sridharan, “Confusability of phonemes grouped according to their viseme classes in noisy environments”, in *Proc. of Australian Int. Conf. on Speech Science & Tech.*, Sydney, 2004, pp. 265–270.

[18] E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem, M. Ozkan, “Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation”, in *Proc. of Signal Proc. and Communications Applications*, 2007, pp. 1–4.