MINISTERUL EDUCAŢIEI NAŢIONALE
ROMÂNIA
UNIVERSITATEA DE MEDICINĂ, FARMACIE, ŞTIINŢE ŞI TEHNOLOGIE „GEORGE EMIL PALADE" DIN TÂRGU MUREŞ

sciendo

Acta Marisiensis. SeriaTechnologica
Vol. 20 (XXXVII) no. 1, 2023
ISSN 2668-4217, ISSN-L 2668-4217

10.2478/amset-2023-0001

ESTIMATION OF THE EXTENT OF VERTICAL HEAD MOVEMENTS FROM SOUND ALONE

Roland KILIK
University of Miskolc
Egyetemvaros, 3515 Miskolc, Hungary
roland.kilik@uni-miskolc.hu

Abstract

Human-like agents are becoming more and more common. However, the usefulness of these agents depends to a large extent on the naturalness of their movements. The classification procedure presented in this article aims to increase the naturalness of the head movements of human-like agents. The method is capable of estimating the vertical range of head movement from the speech sound alone, and thus allows a final phase amplitude correction of the generated head movements of virtual talking heads in order to increase naturalness. The advantage of the method, is that it does not require visual information, works for general subjects, its precision and effectiveness can be improved by defining further classes, and it can improve the naturalness of any head movement generation method's output by a posterior amplitude scaling.

Key words: Head movement estimation, Classification, Virtual agent, Outlier detection, Feature extraction, Human-like agent

1. Introduction

Several studies have shown that a head movement synthesizing system can be made using different sound information as input and little information from visual information as ground truth [1]. Previous research [2] has confirmed that the goal is not to make head movements as close to the original as possible, in the sense that the degree of naturalness of the movement cannot be a measure of similarity to the trajectory of the ground truth. This observation coincides with initial qualitative feedback on neural network synthetized head movements from different subjects [3]. In the course of the initial examination, it was clearly demonstrated that instead of the similarity of the generated head movement's trajectory to the trajectory of the original (recorded) movement, a maximal head movement amplitude which is close to that of the original one, results in a more natural-like impression. This finding led to the research for determining head movement amplitudes from sound information, where estimated amplitude value can be used for automatic amplitude-correction of neural-network-generated head movements. In addition, with the expectation that the solution should not require visual information.

An important publication on automatic motion generation, comparing the results of other authors and their methods' limitations is by Zhou et al. [4]. In this paper, the authors point out that a major problem in generating lifelike motion of virtual talking heads is that the very basic sound information quantities alone are not sufficiently correlated with the motion, making the generation of lifelike (e.g. proper amplitude) head motion based on only sound a difficult task.

The above authors also refer to the work of Hyeongwoo Kim et al. [5], where the authors investigated that a large degradation in quality of motion generation is caused when the estimated/generated motion range is outside the motion range of the training samples of the given movement generation system. This also confirms the importance of estimation of head movement amplitude from the sound (and of course using this in the motion generation process).

The goal of the proposed classification method is to determine the maximal vertical head movement of subjects in spontaneous sentences only from the sound samples. By using the classification, it is possible to reduce the latter problem when generating realistic head movements of virtual talking heads or humanoid robots, using the method to automatically correct the range of the generated movement. In addition to this main advantage, the high correlation value between head motion ranges and the derived voice information quantities used in the classification, can be a starting point for further developments.

The first part of the paper introduces the methodology and describes the classification, which is followed by the application of the introduced method. The third part deals with the results, features, and novelty of the method, followed by the examinations of class-wide correlations and the count of the variables. Finally, a unique outlier detection procedure is introduced whose role is to further improve the efficiency of the classification method.

2. Materials and methods

The chapter describes the collection and analysis of samples as initial steps. After, it presents the details of the classification procedure.

1.1. Sample collection

The goal of the classification method is the classification of vertical head movement amplitudes from sound information, with boundaries set at 15, 25 and 50 pixels. Pixels were chosen as the unit of measurement due to the goal of making the classification possible for a wide variety of subjects/speakers (a large quantity and heterogeneous types of subjects needed to be involved in this study). Therefore, it would not have been impractical to require samples taken in a controlled environment, which is a necessary precondition of estimating angles. Besides this, pixel-based measurement makes it easier to further expand the database for possible improvements.

In order to obtain the samples – taking into consideration the previous statements – YouTube videos were used as input, with heterogeneous (different age, sex) subjects speaking spontaneously in front of a camera. The head movements of whom were even more natural without having markers on their faces.

The samples were extracted from the videos, with each sample being one sentence. It was expected that any body movement made by the speaker while speaking the selected statement should be considerably less (in pixels) than the head movement during the given sentence.

For the establishment of the classification method, 450 samples were used with sound and visual information. Each sample consists of a set of data extracted from one sentence in a given video that satisfies the previously described conditions. Figure 1. shows the video information extraction process for these samples.

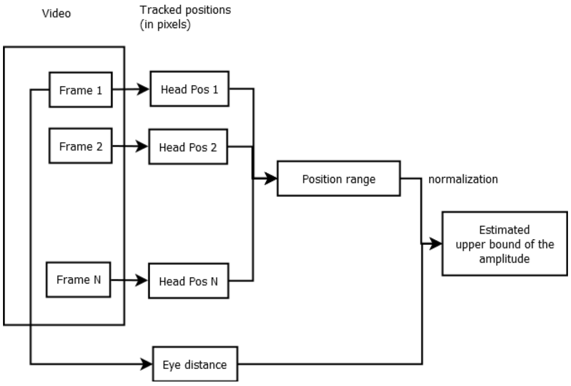

Fig 1. Video information extraction

In addition, a further set of approximately 450 similar characteristic samples (heterogeneous subjects, spontaneous speaking, each sample being one sentence, etc.) were made for testing purposes. For the two sets together the average duration of the sentences was 6.32 seconds. In the whole sample database, no more than 5 samples are produced by one subject in order to avoid sample bias due to any idiosyncratic characteristics of speakers.

1.2. Sample analysis

The maximum of the vertical head movement was given by an eye-tracking method whose results were corrected manually frame by frame. After this step, the movements were scaled to a 50-pixel distance between the eyes, resulting in the distribution of head movement measures that can be seen in Fig 2.

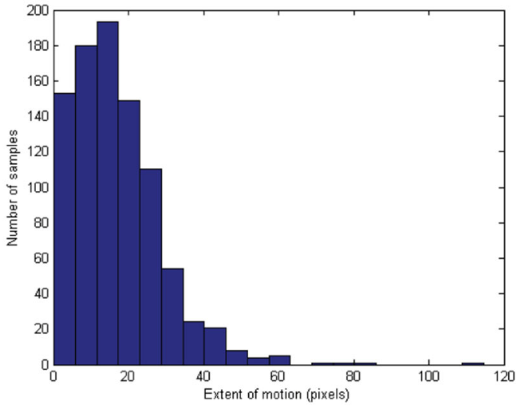

Fig 2. Motion intervals

The sound input of the classification was partly directly extracted values (for example average of intensity, or variance of FFT transformation of the intensity). On the other hand, other sound information values were derived from them on empirical basis. Thus, the total count of the basic and the derived sound information values was 29. These together were used in the class-defining production rules. The extracted properties can be seen in Table 1 in the Appendix. The purpose of the constants is transforming the values to the same interval as others.

Properties P_{14}, P_{15}, and P_{16}were the sample identifier, the measure of movement and the size of the head so these were not used in the production rules, thus from P_1 to P_{32}, altogether 29 sound information properties were represented in the rules of the classes.

1.3. Classification of samples

The classification method partitions the feature space by hyperplanes, which is a well-known approach of the expert systems. We can determine the set of considered points in the feature space by defining production rules. In fact, these rules provide the intersection of the half-spaces expressed by conjunction operators.

With the above mentioned 29 sound information properties, class production rules were made from the whole sample database, where the values of the properties that belong to a class (typically 4-5 per class from the available 29 properties) are in a minimum-maximum range. The samples which are the member of a specific class also have the common feature of their maximal movement amplitude being within a specific domain. Altogether 74 classes were defined with this method (Fig. 3).

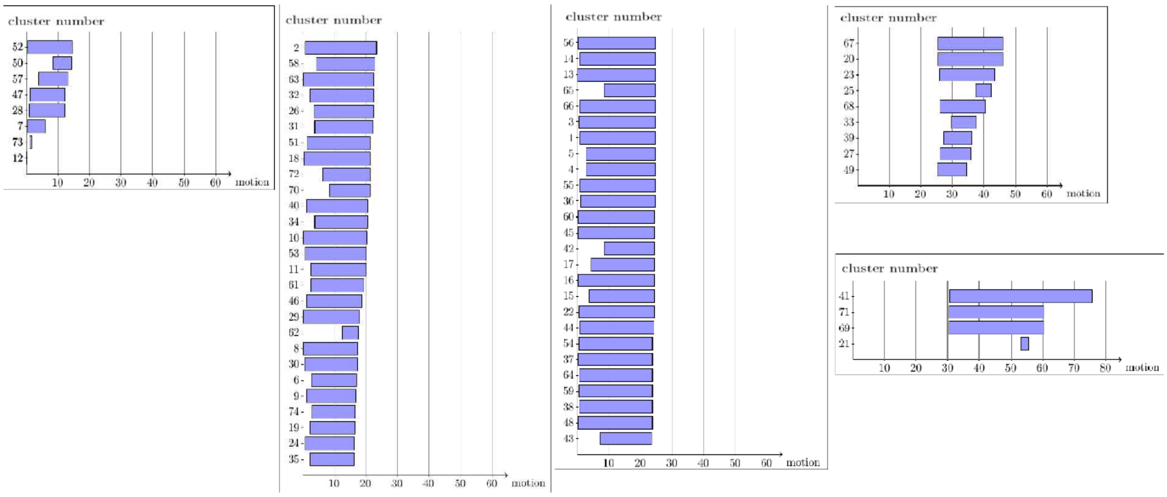

Fig 3. The constructed classes

Let P_j an arbitrary property, $j \in N$. We use comparison operators for defining statements with the corresponding hyperplane. For instance $P_j(,20)$means that the statement is true, when $P_j < 20$. Let define a production rule as a logical formula of these statements, $f: P \rightarrow \{true, false\}$. The production rule also refers to the estimated classes. Therefore, we can formalize the set of our rules as the Rset of (f_i, c_i) pairs, where the f_i is a logical formula on the domain of audio features and c_i is the corresponding class. In the classification process, a sample can belong to multiple classes.

In the next step, for movement amplitudes of 0-15 pixels, 0-25 pixels and higher than 25 pixels, 3 major motion groups were defined containing the previously described 74 classes. The fourth – greater-than-50-pixel – movement amplitude group contained only a few samples at this stage and not included in the study.

If the property values of a sound sample satisfy the rules of any of the minor classes out of the 74, the sample is regarded as part of the class. The fact of belonging to a minor class gives an estimation for the maximal vertical head movement of the sample because every minor class belongs to one of the three large motion groups. With the sum of these, the method at present gives an estimation for the maximum vertical movement in 76 % of the samples.

For example, if the sound information values can be extracted from the sample named $P_1 - P_{32}$ (as can be seen in Table 1 in the Appendix) then if $P_6 \geq 160$, $P_8 < 70$, and $P_4 > 200$, then the sample belongs to class 17 from the 74, which belongs to the major group of movement measuring less than 25 pixels.

The production rules contain "or" and "and" operators in 13 of the 74 classes, and only the "and" operator in the others. The constants of the rule base determined for their adjusted values fulfil the requirement of having the possible minimal number of classification errors.

We regard a sample an outlier when it satisfies the rule of a minor class but is an outlier regarding the movement interval (i.e., by not belonging to the movement measure of the major group that the minor class belongs to). In order to obtain the smallest number of outliers, the values determined by the observation of the data set were machine adjusted. An example of the adjustment can be seen in Fig 4.

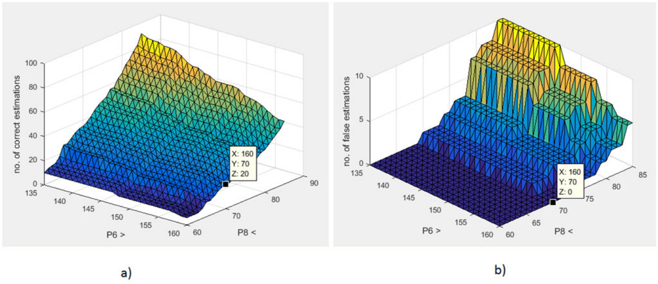

Fig 4. Example of adjusting the constants of a class rule

While Fig. 4a shows the number of samples that are correctly estimated with the actual value of the class properties, Fig. 4b shows the number of misclassified samples with the same property values. It can be seen, that choosing the values to $P_6 \geq 160$, and $P_8 < 70$ is reasonable, because the number of correctly estimated samples reaching the value of 20, while the count of the misclassified samples still remains at the minimum. The adjustment was carried out with the same method for the other 73 classes.

The minor 74 overlapping classes under the three major groups in specific features – bearing in mind that a sample can belong to more than one class – are similar to the entities in the methods of biclustering [6] or coupled two-way clustering [7]. These methods and this one also have a common feature of finding a subset or group of samples with similar characteristics where the group can be described by an interval of a subset of the features available. In our case, a similar characteristic is the movement amplitude in every such group.

The samples belonging to a class that is in the major group of 0-15 pixels are scaled to 7 pixels, while the samples belonging only to a class in the major group of 0-25 pixels are be scaled to 13 pixels (the median of the class-defining samples' movement). The samples in the group of greater-than-25-pixel movement scaled to 36 pixels. If a sample belongs to a class that is in the major group of 0-15 pixels measure of movement and also to a class that belongs to the major group of 0-25 pixels, the more restricted group is the relevant. Figure 5. shows the classification procedure in the form of a high-level diagram.

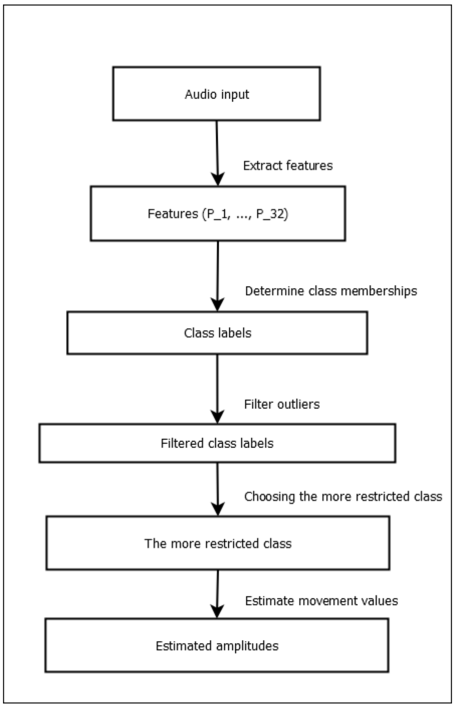

Fig. 5. Classification procedure

3. Results and discussion

This section describes the uniqueness and accuracy of the method, followed by the consideration of high correlation values in the classes. The aforementioned are followed by an outlier detection procedure that improves the method.

2.1. Comparison on accuracy and novelty

Some early head movement synthetizing methods were clustered head motion patterns of 1-2 actors [8]-[10], and Hidden Markov Models were trained for each cluster. However, previous works showed that frame-wise clustering as the first step of the movement generation is not the best approach [11]. The proposed classification solution – as stated above – not using classification in this manner. Any head movement generation procedure can be supplemented with the proposed classification (as a final step), if the generated movement amplitudes can be scaled by a constant value calculated by this classification method.

There have been other attempts (though not with this goal) to produce an implicit estimation for vertical head movement amplitude. The accuracy of it can be calculated from the rate of the generated movement's amplitude and the ground truth [1], [2]. However, their common feature is that the generated head movements were constructed with a few subjects and that the multiplier between the synthesized movement amplitude and the ground truth in a one-sentence interval was 2-10 in the worst case among the presented examples. This ratio can even reach 60 [12], and as the latest result by Matthews et al. [11] can be around 7 according to their example. The previous authors' future goal is seeking a generalization of their method and predict speakers from outside their corpus. Some methods also need ground truth visual information from the beginning of the estimation [2] or from the previous position from the head [9]. Zhou et al. [4] constructed a method for movement generation that is different from the above mentioned ones as it uses only one visual frame, however it focuses only on the lip (and surrounding area) movement generations, and has other limitations. In contrast, the classification method proposed here is constructed for various subjects and does not use visual information for head movement estimations. The proposed method is generalized in the manner that is constructed for various subjects and works for the subjects outside the corpus (except for high emotional cases – the analysis of such cases will be a part of a further study). Its largest estimation error – in case of correct classification – can reach (in the worst case) the 8-10 multiplier, only if the largest amplitude in the sentence is minimal (1-2 pixels) while the sample can only be classified to one of the 0-25 pixel movement interval group's classes and to none of the 0-15 pixel movement interval group's classes. In that case, the amplitude is scaled to 13 pixels (because of the 0-25 pixel group). Situations like this occur in less than 12 % of the samples. Furthermore, the amount can be decreased further by defining more classes in the 0-15 (or even shorter) pixel range – what lowering the chance of a sample with a few pixel motions being only in a 0-25 pixel interval group.

A further result in this area [13] can be attributed to a group of researchers who (as in the previous example) also use mouth movements on the one hand, and individual head movements of selected subjects on the other and generalise from this. In many respects, this leads to a natural solution, but - for example - the large head movements (and possibly other special features) of some subjects in the training samples in some cases lead to unnaturally large head movements in practice. However, this could be corrected by using the present method as an amplitude limiting block.

The rate of the generated and ground truth amplitude in the proposed method is comparable with the latest other researches, while this one not using visual information, generalized regarding the subject, furthermore, the solution is not limited to its current state, but can be enhanced by the definition of new classes.

A smaller difference between estimated and actual amplitude (or head position) than obtained by the presented method could only be achieved by methods that using image information with the speaker (thus with very limited applicability).

For movement generation solutions without visual information, in addition to excessive amplitude, a typical output error is that the method generates an amplitude close to zero for given subjects [4], [14]. Both types of these errors of movement generation can be reduced by automatic amplitude correction using the proposed classification method as an additional phase of the given solution. In addition, it can also be used to improve the output of solutions that use visual information.

In addition to the above, a possible improvement of the method was analysed. If a sample satisfies the rules of more than one class from the 74, we can give the middle of the shortest of those classes' movement intervals as the result. Using this method, the possible movement interval can be further shrunk in 14 % of the samples in the classes that belong to the 0-25 pixel movement interval group, and in 42 % of the samples that belong to a class in the greater-than-25-pixel movement group, with more than 12 pixels in both cases. Although the method adds an additional step to the amplitude scaling, it yields higher accuracy.

2.2. *Class-wide strong correlation and consequence*

In the case of the 42 classes with at least 10 samples, a proportion can be found (for example $P_{27}/P_4/P_{19}$) for the class-defining samples, where the linear correlation between the value of the proportion and the movement is typically over 0.9 for the elements of the class. The average of this correlation value is 0.8758 for the 42 classes and the maximum of it is 0.9979.

For example, let the sound sample with an unknown extent of motion belong to the class where the 0.9 correlation is between the movement and the value of $P_{27}/P_4/P_{19}$. In this case, the value of the proportion and the linear relationship gives a more accurate motion range estimation for that sample than is implied by being part of the class.

In contrast, when examining the same properties after joining every two separate classes of the production system, the correlation value fell between 0.03-0.1 and is typically 0.05 regarding the whole database. This result confirms the validity of the current classification, and of each of the classes within it.

Ben Youssef et al. 2013 [15] have achieved correlations for the low number of subjects with CCA (Canonical Correlation Analysis) around 0.2 between speech and head movement features. Some researchers [12] earlier stated that correlation values only around 0.07-0.08 can be found. Yehia et al. [1] reported strong correlation between intensity and head motion, however only for a low number of subjects, the relationship was varied from utterance to utterance, and the estimation of head motion from the intensity had worse results than the intensity estimation from the movement (avg. 0.37 correlation coefficient). Considering these results, it can be stated that a frame-wise global correlation between speech and motion features is weak.

The previous authors' correlation method (Canonical correlation Analysis) is different from the Pearson correlation that used here, and also the data as the maximal movement amplitude was examined in this study. However, the statement of weak correlation matches with the obtained results, as experiences in this research also showed that the database-wide general connection between basic speech features and motion amplitude is very weak (typically 0.05). However, a class-wide strong correlation is found here, using derived speech information. This confirms the usefulness of the derived sound information features and the classes based on them. Adding that the definition of additional classes would increase the movement amplitude estimation accuracy due to the characteristics of the system, while their merging would obviously reduce it.

2.3. *Examining the count of variables*

In order to possibly reduce the number of variables (that is 29 basic and derived data together), firstly two linear methods (Principal Component Analysis and maximum likelihood-based factor analysis) were examined.

Since factor analysis does not introduce new variables without meaning but makes the interpretation of the original ones possible by their common factors, one possibility for simplifying the production rules is to replace their variables by the common factors. Besides, when writing the variables of the rules of the pairs that are candidates for contraction with factors, possible contractions can be more easily seen.

For this purpose, factor analysis was applied to the classes that contain at least 7 samples. From the typically 4–5 variables per class, the specific variance was under 0.1 for maximum one variable in each class;

thus, a high percentage of the variable's variance can be covered only in one case in every class by this procedure. The minimum of the per class averages of specific variances is 0.29, while the total specific variance average is 0.5367. Thus, reducing the variable count by factor analysis is not reasonable.

With Principal Component Analysis, the variables of the classes could be covered by 1–2 principal components per class. However, it emerges as a disadvantage, that the reconstructed values would have a 25–100 % difference to the original (which would add excessive error into the system), and the procedure also would introduce variables without meaning as principal components. Thus, applying PCA to this system is not reasonable.

In some nonlinear methods – GDA (Generalized Discriminant Analysis) [16], NPE (Neighborhood Preserving Embedding) method [17] and Locally Linear Coordination (LLC) [18] –, it was investigated with their usual parameters whether the movement measure classification could be realized from the sound information by applying k-means or linkage procedures on their output. The number of the dimensions were 2, 4, 6 and 8 for all three procedures. In the case of GDA, the additional parameter was the Gaussian kernel, while it was 12 neighbours for NPE. When using LLC – Locally Linear Coordination – the parameters were 12 analysers and 200 iterations. The evaluation criteria were Gap [19], Davies-Bouldin [20] and Calinszki-Harabasz [21] in all cases. The procedures could not manage to create separate movement measure classes, except for GDA dimension reducing followed by linkage or k-means grouping with silhouette or Davies-Bouldin validation. In these cases, with different GDA dimension numbers, a group formed with more than 100 samples, with movement measure typically under 20. However – in contrast to the proposed solution – this could not produce any classes with other separate movement intervals except the one mentioned, and could not group as many samples.

2.4. *Outlier detection*

A sample is regarded as an outlier in a class if its values satisfy the rule of a given class, but it does not belong to the movement interval of the class. For such samples the system would give a false movement amplitude estimation; that is why their automatic exclusion is reasonable by their features not contained in the class rule. To achieve this goal, and to further improve the above results, the following automatic outlier detection mechanism has been designed.

Defining the rule base of the outlier detection, and exposition of its rules

The classification solution cannot give an estimation in 217 cases from 905 samples at present. Regarding the other currently handled 688 samples, it classifies correctly in 87 percent of the cases and misclassifies in 13 percent (as the movement amplitude of the sample is smaller than 25 pixels but the estimated amplitude is above it, and vice versa). The outliers are among these samples. The above mentioned 13 percent misclassification ratio is without any outlier detection, which was reduced by the following outlier detection procedure. The procedure of the outlier detection system can be seen in Fig. 6.

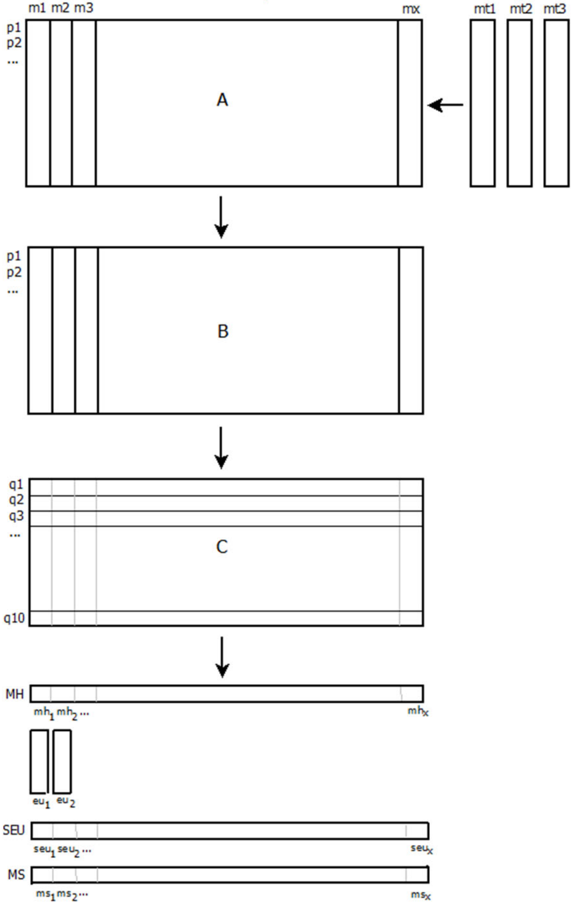

Fig 6. Procedure of the outlier detection

At the first step, we take an *A* class matrix with the elements of $m_1, m_2, \ldots$ We supplement it with the $mt_1, mt_2, \ldots$ outlier detection learning samples, which also satisfy the rule of the class. (From this teaching samples some are outliers, some are not.) Thus, a *B* matrix is constructed from every *A* matrix supplemented with an *mt* sample. In those matrixes, from the $p_1, p_2, \ldots$ sound information values (properties), 10 are chosen in every possible combination, which constructs the matrixes labelled C in Fig. 6 (the names of the actually chosen 10 properties are $q1, q2, \ldots, q10$).

In Fig. 6, which describes the procedure of the outlier detection, $mh_1$is the Mahalanobis distance of $q1{:}q10$ sound information property vector of *m1* column, from matrix C, where m1 is the first sample. Similarly, mh_2 is the same for the second sample, etc. Eu_1 is the Euclidean distance between $q1 \ldots q10$ values of the first column of C matrix (1st sample) and the same *q1:q10* indexed values in the other samples. Eu_2 is the same for the second column of matrix C, etc. Seu_i is $std(eu_i)$ and $ms_i = mh_i / seu_i$.

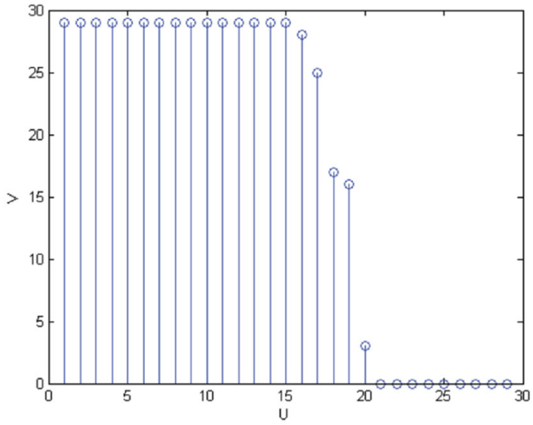

Fig 7. Adjusting outlier detection parameters

In matrix C, the ms_1, ms_2 elements what were obtained from the original class matrix, and the ms_x element what obtained from the actual sample, together produce the vector MS. If the ms_x element's value is the minimum or the maximum in MS, then the given $q1, \dots, q10$ property identifiers are kept. In order for a given $q1, \dots, q10$ property identifier combination (for example $p_1, p_3, p_4, p_8, p_{10}, p_{13,}p_{19}, p_{20}, p_{22}, p_{23}$) to be regarded as valid, there are two conditions. The first is that from multiple classes, it is true for at least 4 outliers that the ms_x element that is calculated for the actual element with the given $q1, \dots, q10$ combination is the minimum or the maximum in the MS vector produced by its own class. The second condition is that the same ms_x element is the minimum or maximum in none of the classes with the same $q1, \dots, q10$ properties, if the sample is not an outlier. If a $q1, \dots, q10$ combination is valid, then for every MS vector that is calculated from this for a future test sample and the base samples of the class, the sample is regarded as an outlier if the ms_x element is minimum or maximum.

It was examined that how many outliers can be covered by the method in the classes with at least 10 elements, using the approximately 450 test samples that were not used in the construction of the class rules. The number of outliers was 29 in the classes of elements with movement amplitude less than 25 pixels, and 23 in the classes of elements with movement amplitude greater than that (thus 29 and 23 were the maximum detectable outliers).

It was also investigated how the results change if combinations containing not 10, but 5, 7, and 12 sound properties chosen, and also the case when for 10 sound information properties the elements of the MS vector are calculated as $ms_i = (mh_i - seu_i)/mh_i$ instead of $ms_i = mh_i/seu_i$.

Fig. 7 shows the optimal case from the *above-25-pixel* movement measure classes regarding the outliers. On the horizontal axis, the number of outliers can be seen that are covered by the total of the combinations (V). The vertical axis shows the numerical value of the requirement (U) that every property combination should occur in at least how many outliers. In the case that can be seen in the figure, combinations were chosen from 5 sound information properties ($q1, \dots, q5$) with $ms_i = mh_i/seu_i$. We can see that even with the requirement – stricter than the original – that every property combination has to occur in at least 15 outliers (and not occur in any of non-outliers), the system can cover all the 29 outliers. For the classes of movement measure under 25 pixels, the best result is also with $ms_i = mh_i/seu_i$, but with 10 element combinations. In this case, however, the maximum number of outliers (here 23) could be covered with the expectation of the combinations occurring in 5 outliers.

4. Conclusions

This paper introduced a method for the classification of vertical head movement amplitude measures from sound information. The advantage of the method is that – in contrast to other approaches – it does not require preliminary movement samples from the subject, and as a result of the nature of its creation, it works for universal subject.

The classification solution at present can give an estimation for the maximal movement amplitude in 76 percent of the samples from the sound alone. This percentage – in contrast to pre-trained methods – can be further increased by defining new classes that contain the uncovered parts of the space, and the accuracy of the estimation can be enhanced by defining smaller movement interval classes. For the cases where the classification method can currently provide estimation, the solution gives a good estimation in 87 percent without any outlier detection. In order to improve this further, a unique outlier detection procedure was also developed and presented that further enhanced this as shown (with the possibility to increase the correct estimation rate to even 100 percent).

The paper described the procedure, the necessary sound information extraction, the method of generating the classes and the properties of the groups, drawing the reader's attention to the high sound-motion correlation values available within the classes, which offer further possibilities. The comparison of the procedure with other solutions, its effectiveness and the aforementioned improvements are also discussed in the paper.

A possible approach to the further improvement of the classification solution presented is classifying to the shortest movement interval class in the case of multiple class membership.

The proposed classification method can improve the naturalness of the outputs of head movement generation methods in a large proportion of cases, by giving a possibility of a final phase amplitude correction based on the classification.

References

[1] Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). *Linking facial animation, head motion and speech acoustics*. Journal of phonetics, 30(3), 555–568.

[2] Greenwood, D., Laycock, & S., Matthews, I. (2017). *Predicting head pose from speech with a conditional variational autoencoder*. Interspeech 2017, 3991-3995.

[3] Czap, L., & Kilik, R. (2015). *Automatic gesture generation*. Production Systems and Information Engineering, 7, 5–14.

[4] Zhou Y.,Han X., Shechtman E., Echevarria j., Kalogerakis E., & Li D. (2020). *MakeltTalk: speaker-aware talking-head animation.* ACM Transactions on Graphics (TOG) 39, 6, 1–15

[5] Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., ... & Theobalt, C. (2018). *Deep video portraits*. ACM Transactions on Graphics (TOG), 37(4), 1-14.

[6] Cheng, Y., & Church, G. M. (2000). *Biclustering of expression data*. Ismb, 8, 93–103.

[7] Getz, G., Levine E., & Domany, E. (2000). *Coupled two-way clustering analysis of gene microarray data*. Proceedings of the National Academy of Sciences, 97(22), 12079–12084.

[8] Deng, Z., Narayanan, S., Busso, C., & Neumann U. (2004). *Audio-based head motion synthesis for avatar-based telepresence systems*. Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence, 24–30.

[9] Grimm, M., & Neumann, U., & Busso, C., Deng Z., & Narayanan S. (2005). *Natural head motion synthesis driven by acoustic prosodic features*. Journal of Visualization and Computer Animation, (3-5), 283–290.

[10] Grimm, M., Neumann, U., Busso, C., Deng, Z., & Narayanan S. (2007). *Rigid head motion in expressive speech animation: Analysis and synthesis*. IEEE Transactions on Audio, Speech, and Language Processing, 3, 1075–1086.

[11] Matthews, I., Laycock S., & Greenwood, D. (2018). *Joint learning of facial expression and head pose from speech*., 15, 2484–2488.

[12] Hofer, G., & Shimodaira, H. (2007). *Automatic head motion prediction from speech data*. In Interspeech 2007, 722-725.

[13] Ji, Xinya, et al. (2022). *Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. ACM SIGGRAPH 2022 Conference Proceedings*. 2022.

[14] Lu, Y., Chai, J., & Cao, X. (2021). *Live speech portraits: real-time photorealistic talking-head animation*. ACM Transactions on Graphics (TOG), 40(6), 1-17.

[15] Ben Youssef, A., Shimodaira, H., & Braude, D. A. (2013). *Articulatory features for speech-driven head motion synthesis*. Proceedings of Interspeech, Lyon, France.

[16] Baudat, G., & Anouar, F. (2000). *Generalized discriminant analysis using a kernel approach*. Neural computation, 12(10), 2385–2404.

[17] Liu, X., Yin, J., Feng, Z., Dong, J., & Wang Lu. (2007). *Orthogonal neighborhood preserving embedding for face recognition*. In Image Processing, 2007. ICIP 2007. IEEE International Conference, 1, 133-136.

[18] Roweis, S. T. et al. (2002). *Automatic alignment of hidden representations*. Sixteenth Annual Conference on Neural Information Processing Systems, Vancouver, Canada, 15, 841–848.

[19] Tibshirani, R., Walther, G., & Hastie T. (2001). *Estimating the number of clusters in a data set via the gap statistic*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411–423.

[20] Davies, L. D., & Bouldin, W. D. (1979). *A cluster separation measure*. IEEE transactions on pattern analysis and machine intelligence, 2, 224– 227.

[21] Calinszki, T., & Harabasz, J. (1974). *A dendrite method for cluster analysis*. Communications in Statistics-theory and Methods, 3(1), 1–27.

Appendix

Table 1 Extracted sound properties

P_1	$\overline{autocorrelation_{pitch}}$
P_2	$var(FFT(intensity))$
P_3	$\frac{\sum_{i=2}^{N}\lvert amp_i - amp_{i-1}\rvert}{No.offrames}$ where: $amp_k = \frac{\sum_{j=k-r}^{k+r}\lvert wave_j\rvert}{2r+1}, r = 11$
P_4	$\frac{\sum_{i=2}^{N}\lvert intensity_i - intensity_{i-1}\rvert}{No.offrames}$
P_5	$\frac{\sum_{i=2}^{N}\lvert pitch_i - pitch_{i-1}\rvert}{No.offrames}$
P_6	$\overline{crosscorrelation_{pitch}}$
P_7	$\frac{P_1 + P_2 + P_5 + P_6 + P_8}{5}$
P_8	$P_4 * 200$
P_9	$var(FFT(amp))$
P_{10}	$var(FFT(pitch))$
P_{11}	$\frac{P_7}{P_8}$
P_{12}	$\frac{P_3 * 100}{P_4}$
P_{13}	$\frac{P_{12}}{P_4}$
…	…
P_{17}	var(amp)
P_{18}	std(amp)
P_{19}	var(intensity)
P_{20}	std(intensity)
P_{21}	$\frac{P_5}{P_{19}} * 1000$
P_{22}	$\frac{P_1}{P_2}$
P_{23}	$\frac{P_4}{P_3}$
P_{24}	$\frac{P_4}{P_5}$
P_{25}	$\frac{P_{11}}{P_2}$
P_{26}	$\frac{P_{26}}{P_{18}}$
P_{27}	$\frac{P_{19} * 10}{P_{22}}$
P_{28}	$\frac{P_{19} * 10}{P_{23}}$
P_{29}	$P_{28} - P_{29,}$
P_{30}	$\frac{P_{28}}{P_{29}}$
P_{31}	$(P_{19}/P_{22})/P_1/P\ 2$
P_{32}	$P_{33} = \frac{P_{23}}{P_{25} * 100}$