

# Multimodal Annotations and Categorization for Political Debates

Brigitte Bigi  
Laboratoire Parole et  
Langage, CNRS &  
Aix-Marseille Universités  
5, avenue Pasteur, BP80975  
13604 Aix en Provence,  
France  
brigitte.bigi@lpl-aix.fr

Cristel Portès  
Laboratoire Parole et  
Langage, CNRS &  
Aix-Marseille Universités  
5, avenue Pasteur, BP80975  
13604 Aix en Provence,  
France  
cristel.portes@lpl-aix.fr

Agnès Steuckardt  
Praxiling, CNRS & Université  
de Montpellier 3  
17, rue Abbé de l'Épée  
34090 Montpellier  
agnes.steuckardt@univ-  
montp3.fr

Marion Tellier  
Laboratoire Parole et  
Langage, CNRS &  
Aix-Marseille Universités  
5, avenue Pasteur, BP80975  
13604 Aix en Provence,  
France  
marion.tellier@lpl-aix.fr

## ABSTRACT

The paper introduces an annotation scheme for a political debate dataset which is mainly in the form of video, and audio annotations. The annotation contains various information ranging from general linguistic to domain specific information. Some are annotated with automatic tools, and some are manually annotated. One of the goals is to use the information to predict the categories of the answers by the speaker to the disruptions. A typology of such answers is proposed and an automatic categorization system based on a multimodal parametrization is successfully performed.

## 1. INTRODUCTION

This work was conducted to analyse political debates in a multimodal perspective. Particularly, this study focused on the answers produced by a main speaker after he was disrupted. The approach relies on the annotations of a debate and on their review. This implies technical and methodological levels to produce high quality multimodal annotations. Then these different types of annotations are merged into a unified framework taking cues from different modalities encoding gestures, prosody, intonation, linguistic information, acoustic information, etc. The aim was to analyse repartee via this multimodal perspective: how MPs can answer to disruptions, what strategies are used, with which modalities and how they are correlated.

Multimodal annotation is often reduced to the annotation of gesture, possibly accompanied by another level of linguistic information (e.g. morpho-syntax). This paper follows a different route. A broad-coverage approach, aiming at annotating a large set of linguistic domains is proposed. This paper reports the methodology used to annotate manually or automatically the corpus (section 2): a video of a political debate at the French National Assembly by Yves Cochet (an ecologist MP). This methodology is based on what was used prior to annotate CID - Corpus of Interactional Data [5]. The annotations and conventions were adapted for the purpose of the specific study of answers to disruptions. A typology of these answers was then proposed (section 4) depending on strategies the MP was using. Finally, an automatic categorization system based on a multimodal parametrization was trained from these data (section 5). Classification scores leave us to consider that the proposed typology is relevant.

Before entering into the detail of the annotations for different linguistic phenomena, it is necessary to underline that the different information domains to be annotated (see section 3) do not rely on the same granularity level and then require a different amount of work. As a consequence, this work is still in progress. However, we already have significant results, as it will be shown hereafter.

## 2. CORPUS DESCRIPTION

The corpus is a political debate at the French National Assembly, on May 4th, 2010. A subset of about 4 minutes was selected. An ecologist MP, Yves Cochet, debates about a law named “Grenelle II de l’environnement” to be voted. Y. Cochet was speaking about electricity, Greenhouse gas emissions, and the electricity production and the pollution they produce. These subjects generated many reactions and MPs disrupted Y. Cochet by remarks or laughs.

The video was downloaded from the National Assembly web

site. It is a standard Flash Video file. The quality is good enough but the definition is small: 320x256. However, it allowed us to annotate gestures, except in some rare cases: sometimes the camera focused on the Assembly instead to focus on the main speaker. The audio track was extracted from the video. The quality is good enough to produce reliable annotations. When MPs in the Assembly are speaking loudly, the audio track contains noises. This hinders the coding of the prosody annotation level but not the other domains, as the main speaker can be heard well enough.

### 3. MULTIMODAL ANNOTATIONS

This section is dedicated to the different tools and conventions used to annotate. It illustrates the kind of process to implement in the perspective of obtaining rich and broad-coverage multimodal annotations.

#### 3.1 Annotation encoding

Multimodal annotation is faced with the necessity of encoding many different information types, from different domains, with different levels of granularity. Actually, the focus is on video corpus with annotations such as prosody, syntax, gestures, etc. Each of these annotations is aligned on the signal and/or the video.

Linguistic annotation, especially when dealing with multiple domains makes use of different tools during the same project. The annotation itself should be done in the most ergonomic and convenient tool for the given annotation task. Then it is up to the person to develop the exploitation tool to create “views” in this very rich and layered object corresponding to the whole annotation set. In this way, analysts, annotators and end-users never face the intrinsic complexity of a rich multi-level annotation framework. Acoustic analysis, phonetics and prosody annotation is very often done using Praat [8]. Gestures, and more generally multimodal studies, now rely on higher level systems such as Anvil [12] or Elan [23]. Not to speak of orthographic transcription which can be done with many tools. None of these tools are directly interoperable, each using a native format, some of them on top of XML, some others developing an ad hoc markup language.

To merge all the annotations, the TextGrid data file format was chosen. It is a Praat native file format based on an ASCII encoding, and many other tools can import it. The limitation of this choice is that TextGrid has a flat representation, and Elan or Anvil can produce tiers in a tree. On the other hand, Praat can deal with point tiers (often used for prosody). All the automatic annotation tools we developed have a read/write API for this format which will be distributed on-demand. The transcription and the prosody annotation were made using Praat. Gestures were annotated with Elan, and tiers were exported (by our API) to the TextGrid format.

The video, the wav file and the TextGrid annotation file are distributed under the terms of the GPL license on the Speech & Language Data Repository web site<sup>1</sup>. Figure 1 shows a screenshot of the annotation tiers.

<sup>1</sup><http://sldr.org/sldr000729/en>

**Table 1: Words and phonetic**

Category	Count
Words	746
Filled pauses	4
Silences	137
Phones	2356
Syllables	1030

### 3.2 Annotation description

#### 3.2.1 IPU segmentation

The audio signal was automatically segmented in IPUs - Inter-Pausal Units. IPUs are blocks of speech bounded by silent pauses of more than 200 ms (this duration is well suited for French), and aligned on the speech signal.

#### 3.2.2 Enriched Orthographic Transcription

The transcription process takes as input the set of IPUs. The corpus was then orthographically transcribed following transcription guidelines in line with the French GARS conventions [7]. This transcription can be seen in the first tier of Figure 1.

One of the characteristics of speech is the important gap existing between a word’s phonological form and its phonetic realisations. Specific realisation due to elision or reduction processes are frequent in speech data. Therefore transcribers were asked to provide an enriched orthographic transcription (EOT), which includes, for example, manually annotated non-standard events such as: mispronunciations, truncated words, some liaisons, elisions, laughs, etc. Some of these specificities have a direct consequence on the phonetisation procedure and so on the syllabification. From the EOT, two transcriptions are generated automatically : (1) the standard orthographic transcription from which the orthographic tokens are extracted to be used for the syntax analysis and its related tools (POS tagger, parser, etc.); (2) a specific transcription from which the phonetic tokens are obtained to be used by the grapheme-phoneme converter. The time spent for this annotation is 1 hour for 1 minute of speech.

#### 3.2.3 Phonetization

An automatic grapheme-phoneme converter was first applied from the EOT. It produced as output a sequence of phonemes coded in Sampa<sup>2</sup>. An automatic alignment was applied to obtain each phoneme time localization. Then, automatic syllabification was performed using the system described in [4]. Finally, from the time aligned phoneme sequence plus the EOT, the orthographic tokens were automatically time-aligned. Table 1 recaps the main figures about the EOT and the phonetization. Figure 1 shows aligned words on tier 2, syllables on tier 3 and phonemes on tier 4.

#### 3.2.4 Prosody

Prosody can be conceived as the linguistic study of rhythm and melody of speech: metrical variation (rhythm) and intonational variation (melody) play a major role in the organization of speech flow by grouping words in different lev-

<sup>2</sup>Speech Assessment Methods Phonetic Alphabet

Figure 1: Annotation example

Table 2: Prosody

Category	Count	Category	Count
RF1	25	RF2	2
RL	5	RT	2
RMC	60	mc	16
F	25	L	1

els of hierarchically structured "meaning groups" [10]. The present study focuses on intonational tunes. These melodic schemas associated with meaning groups differ from one language to another but most languages have paradigms of contrastive tunes. These tunes assume pragmatic roles in conversation, namely by carrying the attitudes of the speakers towards the transmitted contents [3]. Their argumentative role is therefore crucial. Working on French, we have adapted the intonational inventory proposed in [2]. This inventory comprises two categories of minor tunes (a rising and a falling one) whose scope is roughly the phrase, and four categories of major tunes (falling, rising, rising-falling and falling from penultimate) whose scope is rather the sentence or the clause. Three additional functional distinctions are made within the rising tune category: the continuation rise, the list rise and the final rise. The data is annotated through auditive identification realized by professional phoneticians. The time spent for this annotation is about 4 hours for 1 minute of speech. Table 2 recaps the main figures about the intonational tunes annotation. Figure 1 shows "RMC" and "mc" examples in the tier number 5.

### 3.2.5 Gestures

When we code "gestures" we only consider co-speech gestures which normally co-occur with speech. They are movements of the hands and arms produced by people when they talk. They do not belong to a fixed repertoire as gestures of sign language for instance, on the contrary, they are unique, personal and spontaneous. Therefore we do not take into account gestures that do not relate to speech such as scratching one's face, nervous gestures of self touching, etc. There are

several classifications of gestures, McNeill's [16, 17] is currently widely used in the community of gesture researchers even if it sometimes needs to be adapted. The formal model we used for the present annotation of hand gestures is thus adapted from the specification taken from McNeill's work. Iconic gestures bear a close formal relationship to the semantic content of speech [16], they usually illustrate concrete ideas. Metaphoric gestures are very similar to iconics except that they are rather used to depict abstract concepts or illustrate things metaphorically. If one cups their hands when saying the word "pollution" for instance, it is a metaphoric gesture because the cup acts as a symbolic image for the concept of pollution.

Deictics gestures refer to things by pointing with the hand, the finger, the chin, etc. They can be either concrete when pointing to someone, something or somewhere or abstract when referring to something/someone absent or a place or even a moment in time. Finally, beats are rhythmic movements that have no semantic connexion to the speech they accompany. They rather stress important words or phrases. A typical beat would be a flick of the hand or of the finger. Beats are often superimposed on other gestures [17]. To this typology we added two types of gestures: emblems which are conventionalized and cultural gestures with a fixed form such as a thumb up to say "OK" and interactive gestures which are used to influence directly the conversational interaction [1]. We also added a possibility to code aborted gestures (i.e. gestures that begin but are left unfinished).

Table 3 summarizes these annotations. Figure 1 shows a main gesture type "metaphoric" and a "beat" in the tier number 6, without a secondary gesture type (tier 7). All types of gestures have been implemented in Elan with a controlled vocabulary. The corpus presents 122 gestures produced by Yves Cochet. Two tiers was used to code gesture type: number 6 represents the main gesture type (that is where all gestures are coded) and number 7 is called "secondary gestures type" and enables us to mention when a gesture has two dimensions and when there are superimposed beats.

**Table 3: Gesture Types**

Gesture	Main Count	Second. Count
beat	22	50
metaphoric	73	7
deictic	12	7
emblem	7	2
interactive	2	5
iconic	4	
aborted	2	

Finally, handedness was coded (tier number 8) since in this particular discourse it seems relevant. In case of a single-handed gesture, we coded it in its “Handedness”: left or right hand. In case of a two-handed gesture, we coded it as if both hands moved in a symmetric way or if the two hands moved in an asymmetric way (Table 4). The time spent for gesture annotation is about 2 hours for 1 minute of video.

**Table 4: Handedness**

Hand	Count
left	72
right	26
both, symmetric	23
both, asymmetric	1

### 3.2.6 Syntax

The stochastic parser used to annotate syntax is described in [6]. It generates automatically morpho-syntactic and syntactic annotations. The parser has been adapted in order to account for the specificities of speech analysis. First, the system implements a segmentation technique, identifying large syntactic units that can be considered as the equivalent of sentences in written texts. A second modification concerns the lexical frequencies used by the parser model in order to capture phenomena proper to speech data. The categories and groups counts for the whole corpus are summarized in Table 5. Figure 1 shows categories on tier 9 and groups on tier 10 (GA: adjective phrase; GN: noun phrase; GP: prepositional phrase; GR: adverbial phrase; NV: verb nucleus; PV: group containing a preposition introducing a verb).

### 3.2.7 Self-repetitions

Self-repetitions were manually annotated. The source and the repetition were categorized as: DIS (disfluent) ; RHE

**Table 5: Syntax**

Category	Count	Group	Count
adverb	94	GA	42
adjective	52	GN	101
auxiliary	4	GP	68
determiner	77	GR	84
conjunction	58	NV	110
interjection	9	PV	4
preposition	77		
pronoun	110		
noun	136		
verb	107		

**Table 6: Repetitions**

	Source Count	Repetition Count
DIS	16	12
RHE	27	23
LIST	8	6

(rhetoric) ; LIST (list). Counts are presented in Table 6. Sometimes, there are several sources finalized by the repeat. This is the reason why the number of sources is higher than the number of repetitions. This was about 45 minutes annotation for 1 minute of speech. Figure 1 shows repetitions on tier 11.

## 4. ANSWERS TO DISRUPTIONS

This annotation consists in the proposition of a typology based on a manual analysis of the debate report. This report includes remarks of MPs in the Assembly that can not be heard in the audio track. This enabled us to align on the signal all parts of the discourse which are concerned by the answers. Figure 1 shows this annotation on tier 12.

### 4.1 The contempt strategy

This strategy is used twice in the selected part of the debate. It consists in ignoring the disruption and continuing the discourse. This strategy can be observed in the video with some converging evidence factors: a short silence, an head movement going to the right (where the disruption come from) and coming back to the front, some specific face features (that we interpret as contempt). These disruptions are mentioned in the report of the debate by “laughts” or by the writing statement of the MP who caused the disruption.

### 4.2 The direct answer

This strategy is divided into two sub-categories. In the first, Yves Cochet just says directly “no” (or another short negative answer), without adding any arguments. The second is made of two steps: Yves Cochet repeats the phrase (which is an other-repetition), and then he argues. Figure 2 illustrates a direct answer<sup>3</sup> with the use of irony. In this case, the main gesture is deictic and coupled with an interactive, using right hand. The intonational tune is a falling.

### 4.3 The “sideway answer”

This strategy does not consist in answering but in commenting the disruption. This comment can be directed to the Assembly (meta-discourse) or it can be addressed to the chair (meta-interaction) [22]. Figure 3 illustrates this latest case. Yves Cochet turns to the chair and he uses low tones.

## 5. AUTOMATIC CATEGORIZATION

A system that performs categorization aims at assigning appropriate categories from a predefined classification scheme to incoming data. Categorization is an important component of many large Information Retrieval or Machine Learning system. It is often defined as the content-based assignment of one or more predefined categories to new observed data. Machine Learning approaches to classification (categorization is a classification task) suggest that the construction

<sup>3</sup>no planes is the Icelandic volcano



Figure 2: Yves Cochet - “[plus d’avions c’est le volcan islandais]”



Figure 3: Yves Cochet towards the chair

of categorization means using induction over pre-classified samples. They have been rather successfully applied in various studies.

This section explores and identifies the benefits of the use of automatic categorization. We suppose that if the proposed typology is consistent with the multimodal annotations, an automatic classification system will be able to model it and to produce correct predictions. Unlike manual annotations that can be subjective, the advantage of the automatic method is that it is objective. The C4.5 algorithm (without pruning) proposed in [19] and implemented in the Weka software [11] was chosen. This choice was motivated by work reported in [14] which indicates that parameter tuning is often more important than the choice of algorithm. Experiments were carried out with a 100 fold cross-validation process on the annotated corpus presented in this paper.

One of the difficulties related to the multimodal aspect is that annotations are *heterogeneous and unsynchronized*. The second difficulty is related to missing values, as they need

to be modeled. Another difficulty is that we can not assess the degree of confidence of the annotations, we then suppose that we are dealing with uncertain data. Data uncertainty is a recent challenge in classification [9, 20, 18, 15]. However, some of these studies simulate uncertainty and all of them are concerned with only one modality.

In order to make a category decision, a representation of categories must be established. One difficulty is to transform annotations, which typically are strings of characters, into a more efficient representation for the learning algorithm and the classification task. Moreover, to cope with the unsynchronized aspect, we oversampled the data as proposed in [24]. One sample is generated for each of the smallest duration annotation: one sample per phoneme in our case. The reference-time is the middle of each phoneme. Then we obtained about 2,500 samples. At a given time  $t$ , each modality is represented as a vector over the set of possible annotations for this modality. The construction of a categorization classifier for category  $c_i \in C$  consists in the definition of a function that, given a sample  $s_t$ ,  $t \in T$  returns a categorization status value for it. All these vectors are concatenated to obtain a classification vector for each sample: it is an early fusion process as proposed in [13]. Let  $s_t^m$  be the sample of the modality  $m$  at time  $t$ . This sample is represented by a vector containing annotations as:  $s_t^m = (a_t^{m,1}, a_t^{m,2}, \dots, a_t^{m,n}, \dots, a_t^{m,N^m})$  where  $a_t^{m,n} = 1$  if the annotation is affected to the sample for the modality  $m$  and  $a_t^{m,n} = 0$  otherwise. Then, the multimodal vector  $\vec{s}_t$  of a sample  $s_t$  is defined as:  $\vec{s}_t = (s_t^1, s_t^2, \dots, s_t^m, \dots, s_t^M)$ . Experiments using this system produce 90.63% of correct classifications.

This first system produced encouraging results, but it does not differentiate empty annotations and missing annotations. In our data, 117 samples are missing for the prosody annotation and 74 samples are missing for the 3 gestures tiers. The pruning solution implemented in Weka was tested and improved the classification score to 91.08% which is unsatisfactory (the relative gain is of 4.80%). Various pruning solutions for the C4.5 algorithm was compared in [21] and results was more or less similar for all methods. We adopted an easier way to take into account missing values and we fixed equal values for each missing annotation as:  $a_t^{m,n} = \frac{1}{|N^m|}$ . This system produced 94.33% of correct classification, which is a significant improvement (the relative gain compared to the first system is 39.49%).

To deal with uncertainty of multimodal annotations, we propose to assign an  $\epsilon$  probability to non-observed annotations. In a multimodal context, an  $\epsilon_m$  for each annotation in a modality  $m$  was manually assigned. A small  $\epsilon_m$  values was used to manual annotation ( $\epsilon_m = 0.001$ ) and a bigger one to automatic annotated domains ( $\epsilon_m = 0.015$ ). Then if a mass is attributed to unobserved annotations, observed annotations must be pruned. Consequently,  $s_t^m$  is made of annotations estimated as:  $a_t^{m,n} = \epsilon_m$  if this annotation is not assigned to this sample and  $a_t^{m,n} = a_t^{m,n} - \lambda_m \epsilon_m$  otherwise.  $\lambda_m$  is the number of non-observed annotations for the modality  $m$  of the sample  $s_t$ . This system produced 94.98% of correct classification. The relative gain compared to the previous system is 11.46%, which is a significant.

Correct classification scores for each individual modality<sup>4</sup> are interesting: prosody only is 73.83%; gesture types only (main and secondary) is 79.14%; handedness only is 78.38%; syntax only is 72.91% and repetitions only is 72.71%. Each of these annotations improved the classification score in the multimodal system.

The final high classification score leads us to consider that the proposed typology is relevant because the multimodal classifier is able to model and predict the proposed categories with a high degree of confidence. This confirms that 1/ multimodality is a good way to analyse this kind of data and 2/ annotations are a key point in this field. Particularly, gesture and prosody annotations are the most important.

## 6. CONCLUSION

A methodological approach to annotate multimodal data in multiple domains was proposed in this paper. This annotation process was applied to a political debate. The video was recorded at the French National Assembly during a debate about ecology. The analysis of the debate report leads us to propose a typology of the strategies used to answer to disruptions. We trained a specific classification system from the multimodal annotations to validate the typology and validated the approach. Our perspective is to annotate the rest of the video. We plan to use the classification system to annotate the answers to disruptions and to manually validate this annotation.

## 7. REFERENCES

- [1] J.-B. Bavelas, N. Chovil, L. Coates, and L. Roe. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, 21:394–405, 1995.
- [2] R. Bertrand, C. Portes, and F. Sabio. Distribution syntaxique, discursive et interactionnelle des contours intonatifs du français dans un corpus de conversation. *Travaux neuchâtelois de linguistique*, 47:59–77, 2007.
- [3] C. Beyssade and J.-M. Marandin. French intonation and attitude attribution. In *Texas Linguistics Society Conference: Issues at the Semantics-Pragmatics Interface*, Somerville, MA, 2007. Denis et al. (eds.), Cascadilla Press.
- [4] B. Bigi, C. Meunier, I. Nesterenko, and R. Bertrand. Automatic detection of syllable boundaries in spontaneous speech. In *Language Resource and Evaluation Conference*, La Valetta, Malte, 2010.
- [5] P. Blache, R. Bertrand, B. Bigi, E. Bruno, E. Cela, R. Espesser, G. Ferré, M. Guardiola, D. Hirst, E.-P. Magro, J.-C. Martin, C. Meunier, M.-A. Morel, E. Murisasco, I. Nesterenko, P. Nocera, B. Pallaud, L. Prévot, B. Priego-Valverde, J. Seinturier, N. Tan, M. Tellier, and S. Rauzy. Multimodal annotation of conversational data. In *The Fourth Linguistic Annotation Workshop*, Uppsala, Suède, 2010.
- [6] P. Blache and S. Rauzy. Influence de la qualité de l'étiquetage sur le chunking : une corrélation dépendant de la taille des chunks. In *Actes de TALN*, pages 290–299, Avignon, 2008.
- [7] C. Blanche-Benveniste and C. Jeanjean. *Le français parlé*. Transcription et édition, Didier Erudition, 1987.
- [8] P. Boersma and D. Weenink. Praat: doing phonetics by computer, <http://www.praat.org>, 2011.
- [9] M. Chau, R. Cheng, and B. Kao. Uncertain data mining: A new research direction. In *Workshop on the Sciences of the Artificial*, Hualien, Taiwan, 2005.
- [10] A. Di Cristo. Le cadre accentuel du français contemporain. *Langues*, 3(2), 184–205, *Langues*, 4(2), 258–267, 1999.
- [11] G. Holmes, A. Donkin, and I.-H. Witten. Weka: a machine learning workbench. In *Second Australian and New Zealand Conference on Intelligent Information Systems*, pages 357–361. Intelligent Information Systems, 1994.
- [12] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *7th European Conference on Speech Communication and Technology*, pages 1367–1370, Scandinavia, 2001.
- [13] M. La Cascia, S. Sethi, and S. Sclaroff. Combining textual and visual cues for content-based image retrieval on the world wide web. In *IEEE Workshop on Content - Based Access of Image and Video Libraries*, Washington, DC, USA, 1998.
- [14] N. Lavesson and P. Davidsson. Quantifying the impact of learning algorithm parameter tuning. In *The Twenty-First National Conference on Artificial Intelligence*, Boston, USA, 2006.
- [15] C. Liang, Y. Zhang, and Q. Song. Decision tree for dynamic and uncertain data streams. In *2nd Asian Conference on Machine Learning*, volume 3, pages 209–224, Tokyo, Japon, 2010.
- [16] D. McNeill. *Hand and Mind: What gestures reveal about thought*. Chicago : The University of Chicago Press, 1992.
- [17] D. McNeill. *Gesture & thought*. Chicago: The University of Chicago Press, 2005.
- [18] B. Qin, Y. Xia, R. Sathyesh, S. Prabhakar, and Y. Tu. urule: A rule-based classification system for uncertain data. In *10th IEEE International Conference on Data Mining Workshops*, pages 1415–1418, Sydney, Australia, 2010.
- [19] J.-R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman ed., 1993.
- [20] J. Ren, S.-D. Lee, X. Chen, B. Kao, R. Cheng, and D.-W.-L. Cheung. Naive bayes classification of uncertain data. In *Ninth IEEE International Conference on Data Mining*, pages 944–949, 2009.
- [21] M. Saar-Tsechansky and F. Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 8:1625–1657, 2007.
- [22] V. Traverso. *Réplique*, page 502. Dictionnaire d'analyse de discours. Patrick Charaudeau et Dominique Maingueneau (éds), Paris : Seuil, 2002.
- [23] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research, 2006.
- [24] Q. Zhi, M. N. Kaynak, K. Sengupta, A.-D. Cheok, and C.-C. Ko. HMM modeling for audio-visual speech recognition. In *IEEE International Conference on Multimedia and Expo (ICME'01)*, pages 136–139, Los Alamitos, CA, USA, 2001. IEEE Computer Society.

<sup>4</sup>The minimum score is 72.71% by assigning each sample to the most frequent class