Acoustic Emotion Recognition using Dynamic Bayesian Networks and Multi-Space Distributions

R. Barra-Chicote[1], *F. Fernandez*[1], *S. Lutfi* [1], *J.M. Lucas-Cuesta* [1], *J. Macias-Guarasa*[2], *J.M. Montero*[1], *R. San-Segundo* [1], *J.M. Pardo* [1]

[1]Universidad Politecnica de Madrid, Spain
[2]University of Alcala, Spain
barra@die.upm.es[1], macias@depeca.uah.es[2]

Abstract

In this paper we describe the acoustic emotion recognition system built at the Speech Technology Group of the Universidad Politecnica de Madrid (Spain) to participate in the *INTERSPEECH 2009 Emotion Challenge*. Our proposal is based on the use of a Dynamic Bayesian Network (DBN) to deal with the temporal modelling of the emotional speech information. The selected features (MFCC, F0, Energy and their variants) are modelled as different streams, and the F0 related ones are integrated under a Multi Space Distribution (MSD) framework, to properly model its dual nature (voiced/unvoiced). Experimental evaluation on the challenge test set, show a 67.06% and 38.24% of unweighted recall for the 2 and 5-classes tasks respectively. In the 2-class case, we achieve similar results compared with the baseline, with 8.5 times less features. In the 5-class case, we achieve a statistically significant 6.5% relative improvement.
Index Terms: automatic emotion recognition, multi-space probability distribution, dynamic bayesian networks, emotion challenge

1. Introduction

The *INTERSPEECH 2009 Emotion Challenge* [1] has been designed to provide a common platform for experimentation (corpus, features, baseline results), to advance in the field of realistic emotional speech identification.

The challenge addresses three sub-challenges in two different degrees of difficulty by using non-prototypical five or two emotion classes (including a garbage model): The Open Performance Sub-Challenge, the Classifier Sub-Challenge and the Features Sub-Challenge.

In our proposal we only participate in the first one, in which we propose both our own feature set, and the classification strategies to be used.

The proposed tasks have a high degree of complexity, mainly due to two factors:

- The emotional speech was recorded in real conditions (children speech playing with an electronic pet in a realistic environment), in contrast with the vast majority of previous research work, in which emotions were acted, usually by professional speakers.
- The emotional labels are not primary, so that the cultural and sociological background of the speakers may have an impact on the actual emotion generation process, which should be taken into account.

In our system design and definition process we based our decisions in the following items:

- Our feature set will be composed of segmental and prosodic ones, as both of them have proved to be relevant in the identification of given emotions [2], and we do not have accurate speech tools (speech recognition, speaker identification and adaptation, etc.) in the German language.
- Mechanisms to account for the dynamic (temporal) behaviour of the emotional information should be included
- Mechanisms to account for the dual nature of prosodic information (voiced/unvoiced speech) should be modelled.

2. Corpus description

In the challenge, the FAU Aibo Emotion Corpus of spontaneous, emotionally coloured speech is used. It is composed of nine hours of speech (51 children), recorded at two different schools. The corpus further provides a uniquely detailed transcription of spoken content with word boundaries, nonlinguistic vocalisations, emotion labels, units of analysis, etc. Detailed information on the corpus can be found at [12].

3. System Architecture

In this section we describe the feature set used, and the proposed system architecture for emotional speech identification.

3.1. Acoustic Features

It is well known that both vocal tract and source configuration are heavily affected by emotions.

Previous experiments carried out in our Group regarding acted emotional speech identification in Spanish [3][4] and German [5] gave us additional quantitative and qualitative information on the relative importance of acoustic features when dealing with different types of emotions. These works suggest that some emotions mainly provoke significant vocal tract modifications, while others show clearly prosodic modification patterns, so that high identification rates could be obtained with the fusion of both sources of information [2].

Based on these previous experiments we decided to use the first 12 $MFCC$ and their first and second derivatives ($\Delta MFCC$ and $\Delta\Delta MFCC$ respectively) as the features representative of the vocal tract.

Prosody has to do with intonation, rhythm an intensity information. An accurate modelling of the emotion rhythm is difficult due to the unavailability of precise phonetic transcription of the corpus, so that in our system we decided to use intona-

tion and intensity related information only. The intonation related features we used are the logarithm of the fundamental frequency ($logF0$) and its first and second derivatives ($\Delta logF0$ and $\Delta\Delta logF0$ respectively). Regarding intensity, the energy ($logE$) and its derivatives ($\Delta logE$ and $\Delta\Delta logE$ respectively) are modelled too.

Additionally, in order to avoid the F0 speaker bias, F0 information is normalised using the following approach:

- For the training set, F0 information is normalised (using mean and variance) in a per-speaker basis, given that we could find out the identity of the speakers.
- For the test set, F0 information is normalised using a global value for mean and variance, extracted from the whole training set.

The feature extraction process has been carried out using the HTK front-end [6] and the F0 extraction tools provided by HTS [7].

3.2. Emotion Dynamic Modelling

During system design, both traditional HMMs [8] and the more general Graphical Models [9] were considered. Our final implementation makes use of Graphical Models, as they offer a more consistent framework to be able to cope with the combination of different information streams, of different nature (discrete, continuous, etc.).

Additionally, and in order to account for the dual nature of prosodic information (voiced/unvoiced speech) we decided to use a multi-space distribution (MSD) approach, such as in [10]. The MSD strategy can be efficiently implemented using Graphical Models, by allowing the inclusion of explicit dependency relationships among the used variables. More specifically, we are using a Dynamic Bayesian Network (DBN) to integrate all sources of information.

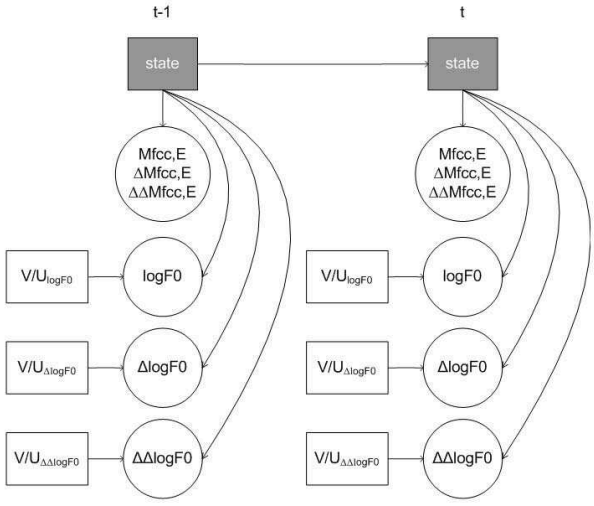

Figure 1: DBN topology

Figure 1 shows the architecture for the Dynamic Bayesian Network used, in which the rectangular boxes represent discrete variables and the circular ones refer to continuous variables. Shaded rectangles represent hidden variables.

The discrete variables are:

- $[state]$: Which models the system state at time t, and system state at time $t-1$.
- $[V/U_{logF0}]$ $[V/U_{\Delta logF0}]$ $[V/U_{\Delta\Delta logF0}]$: Three switching parent variables, each of them deciding on the dependence (or not) between the associated continuous variable and the corresponding parent state variable. If the switching parent variable is considered *V*oiced, the relationship between the parent state variable and the child continuous variable is established. Otherwise, the relationship is not considered. The switching parent variables are consistently calculated taking into account the *voiceness* of the processed frames and the *border effects* in the calculation of the Δ features (this is the reason for the existence of three V/U_* variables instead of only one).

The continuous variables are:

- $(MFCC, E\ \ \Delta MFCC, E\ \ \Delta\Delta MFCC, E)$: A 39-dimensional variable, modelling the behaviour of MFCC and energy related features using standard diagonal covariance Gaussian Mixture Models (GMMs).
- $(logF0)$ $(\Delta logF0)$ $(\Delta\Delta logF0)$: Three variables, each of them modelling the behaviour of the given F0 related feature, also using diagonal covariance GMMs.

The GMTK Toolkit [11] has been used to implement the DBN classifier.

4. Experimental Setup and System Evaluation

For the five-class classification problem, the cover classes are **A**nger (subsuming angry, touchy, and reprimanding) **E**mphatic, **N**eutral, **P**ositive (subsuming motherese and joyful), and **R**est. In the two-class problem, the cover classes are **NEG**ative (subsuming angry, touchy, reprimanding, and emphatic) and **IDL**e (consisting of all nonnegative states).

During the system development process, we only made use of the *challenge train set* provided by the organisation. For fine-tuning and cross-validation purposes, this set has been divided into three subsets, referred as train, devel and test subsets. A detailed description of these partitions is shown in Table 1 and Table 2.

Table 1: *Number of instances for the 5-class task and the corresponding partitions sets for the system training, development and testing.*

	A	E	N	P	R	$\sum$
train	641	1,448	3,951	462	498	7,000
devel	119	314	803	98	125	1,459
test	121	331	836	114	98	1,500
$\sum$	881	2,093	5,590	674	621	9,959

During the training stage, emotion dependent models (DBNs) are generated using the train subset. The number of gaussians has been set according the *Unweighted Identification Recall* (which is the target metric that will be used in the challenge), obtained in the devel subset in order to avoid over-fitting.

During the evaluation stage, for a given utterance, the system accumulates the frame scores for each model, and finally selects the emotion with the best score.

Table 2: *Number of instances for the 2-class task and the corresponding partitions sets for the system training, development and testing.*

	NEG	IDL	$\sum$
train	2,346	4,644	7,000
devel	495	964	1,459
test	517	983	1,500
$\sum$	3,358	6,601	9,959

4.1. MSD validation

During system development we wanted to evaluate the contribution of the MSD modelling to the system performance. In order to do so, we built a baseline system using only the $MFCC$-related features and $logF0$, and applied it to the two-class problem.

When not using the MSD approach, there was no difference in the treatment of voiced and unvoiced frames. The results are shown in Table 3. As it can be clearly seen, the relative improvement of MSD approach is over 10.2%.

Table 3: *Evaluation of the incorporation of MSD into the DBN for the 2-class problem*

	Baseline	Baseline + MSD
Recall	59%	65%
Precision	60%	65%
F	59%	65%

4.2. System Evaluation on the Internal Test Set

The final results of the proposed system, evaluated on the test subset (extracted from the *challenge train set* as described in section 4) are shown in Table 4 (providing the unweighted Recall, the Precision and balanced F-measure). These results are in line with those provided by the organisers in [1], although in a different data set, making useless any further discussion on this comparison.

Table 4: *Evaluation results on the Internal Test Set.*

	2 classes	5 classes
Recall	68%	31%
Precision	69%	33%
F	69%	32%

4.3. System Evaluation on the Challenge Test Set

Tables 5 and 6 show the confusion matrices of the proposed system on the 2-class and 5-class tasks, respectively.

Regarding the 2-class problem, it can be seen that the NEG class is much better identify than IDL class. This is a somehow expected result, as the NEG class is supposed to include speech data pertaining to more specific emotions than the IDL class, which comprises a wider range of emotional content. This result also correlates with the information provided in Table 6 for the 5-class problem, in which it can be seen that the best identified emotion classes are **A**nger (34.5%, subsuming angry, touchy, and reprimanding) and **E**mphatic (77.3%).

Regarding the 5-class problem, it can also be observed that the unbalanced distribution of the data implies a high confusability of all generated emotions with the two classes with more data available (E and N).

Table 5: *Confusion matrix of the 2-class problem*

	NEG	IDL
NEG	79%	21%
IDL	45%	55%

Table 6: *Confusion matrix of the 5-class problem (ref=actual emotion, hyp=recognized emotion).*

ref/hyp	A	E	N	P	R
A	34.5%	48.1%	6.2%	1.1%	10.0%
E	8.6%	77.3%	8.4%	0.4%	5.2%
N	7.5%	49.5%	26.9%	4.6%	11.4%
P	7.4%	18.1%	26.0%	25.1%	23.3%
R	12.3%	30.0%	22.9%	7.5%	27.3%

Table 7: *Comparison between the challenge baseline (using dynamic modelling) and our proposed system (GTH). Relative improvement (RI) is included.*

Task	Baseline	GTH	%RI
2-class	$66.1 \pm 1.02\%$	$67.06 \pm 1.01\%$	**1.5%**
5-class	$35.9 \pm 1.03\%$	$38.24 \pm 1.05\%$	**6.5%**

Table 8: *Comparison between the challenge baseline (using static modelling) and our proposed system (GTH). Relative improvement (RI) is included.*

Task	Baseline	GTH	%RI
2-class	$67.7 \pm 1.01\%$	$67.06 \pm 1.01\%$	-0.9%
5-class	$38.2 \pm 1.05\%$	$38.24 \pm 1.05\%$	**0.1%**

Finally, the baseline results for this dataset provided in [1] are compared with those obtained by our proposed system.

Table 7 compares our results with the best ones provided in Table 4 of [1] employing the low-level descriptors and dynamic modelling. Although both systems are comparable in nature (dynamic modelling and use of standard features related to MFCC, F0 and energy), our system achieves a relative improvement of 1.5% and 6.5% in the 2 and 5-class tasks, respectively. The confidence intervals calculated in Table 7 (for a significance level of 95%) show that only the differences found for the 5-class task are statistically significant.

Table 8 compares our results with the best ones provided in Table 5 of [1] employing the static modelling. In this case, the differences between both systems are minimal and statistically not significant. This is specially relevant since the number of features we are using is much smaller (45 vs. 384).

5. Conclusions

This work describes the implementation of an automatic acoustic emotion recognition system for the *INTERSPEECH 2009*

Emotion Challenge. The implemented system is based on a Dynamic Bayesian Network (DBN) operating on four data streams.

The feature set is composed of 45 elements comprising MFCC, F0 and energy related features, including their first and second derivatives. F0 speaker bias is subtracted in order to avoid speaker dependencies on this feature.

The dual nature of F0 related features (for voiced and unvoiced frames) is modelled using a Multi Space Distribution (MSD) approach, which has been validated showing a relative improvement of 10% when MSD is applied.

The DBN based system results are:

- For the 2-class task, the unweighted recall is 67.06%, which is within the confidence intervals in the comparison to the baseline results provided by the challenge organisers in [1].
- For the 5-class task, the unweighted recall is 38.24%, which significantly improves (6.5% relative) the baseline results provided using also a dynamic modelling strategy. When comparing this result with the one provided using static modelling, the differences are not significative, although in our case we use 8.5 times less features.

6. Future Work

In order to be able to cope with the complexity of realistic human emotional speech identification, it would be necessary to combine acoustic, linguistic and supralinguistic knowledge. However the acquisition of each knowledge source may have important difficulties.

In our case we do not have accurate German acoustic and language models, that would allow us to make use of this extra information (speech recognition, speaker identification and adaptation, etc.). This is the reason why our system is based in acoustic features alone. However, the proposed architecture is flexible enough to be able to include this additional information at a low cost, with a clear foreseeable improvement in the achieved results. So, one of our main interest in future work is getting this additional knowledge sources to be integrated in out system.

Our future plans also include further explotation of the multi-stream capabilities of the DBN model along with the MSD approach, making special emphasis on detailed studies on the impact of speaker-related dependencies and normalization strategies to deal with it.

7. Acknowledgements

This work has been partially supported by project SDTEAM-UPM (TIN2008-06856-C05-03), SDTEAM-UAH (TIN2008-06856-C05-05) and ROBONAUTA (DPI2007-66846-c02-02).

We would also like to thank all the people at the Speech Technology Group for fruitful discussions and for making the implementation of this system possible.

8. References

[1] Schuller, B., Steidl, S. and Batliner, A. “The Interspeech 2009 Emotion Challenge”, Interspeech (2009), ISCA, Brighton, UK, 2009.

[2] Barra-Chicote, R., Montero, J.M, Macias-Guarasa J., D’haro, L.F., San-Segundo, R., Cordoba, R. “Prosodic and segmental rubrics in emotion identification”. In Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1085-1088, May 2006.

[3] Barra-Chicote, R., Montero, J.M., Macias-Guarasa J., Lufti, S., Lucas, J.M., Fernandez, F., D’haro, L.F., San-Segundo, R., Ferreiros, J., Cordoba, R., and Pardo J.M.. “Spanish Expressive Voices: Corpus for Emotion Research in Spanish”. In Proc. of 6th International Conference on Language Resources and Evaluation (LREC2008), May 2008.

[4] Lutfi S.L., Montero, J.M., Barra-Chicote, R., Lucas-Cuesta, J.M. and Gallardo-Antoln, A. “Expressive speech identifications based on Hidden Markov Model”. Proceedings of the International Conference on Health Informatics (HEALTHINF), pp. 488-494. Porto (Portugal). January 14-17 2009.

[5] Barra-Chicote, R, Macias-Guarasa, J., Montero, J.M., Rincon, C., Fernandez, F. and Cordoba. R. “In Search of Primary Rubrics for Language Independent Emotional Speech Identification”. In Proc. of IEEE Workshop on Intelligent Speech Processing, October 2007.

[6] Young, S.J., Evermann, G., Gales, M.J.F., Kershaw, D., Moore, G., Odell, J.J., Ollason, D.G., Povey, D., Valtchev, V. and Woodland, P.C. “The HTK book version 3.4 Manual”. Cambridge University Engineering Department, Cambridge, UK.

[7] The HTS working group, “HMM-based Speech Synthesis System (HTS), http://hts.sp.nitech.ac.jp”. Last access: April 2008.

[8] Gales, Mark and Young, Steve “The application of hidden Markov models in speech recognition”. Foundations and Trends in Signal Processing Vol. 1, No. 3 (2007) 195304

[9] Murphy K.P. “Dynamic Bayesian Networks: Representation, Inference and Learning” PhD. Thesis. California University, Berkeley, 2002.

[10] Tokuda, K., Masuko, T., Miyazaki N. and Kobayashi T. “Multi-Space Probability Distribution HMM”, IECI TRANS. INF. & SYSTEMS, vol. E85-D, n 3, March 2002.

[11] Bilmes, Jeff. “http://ssli.ee.washington.edu/ bilmes/gmtk”. Last access: April 2009.

[12] Steidl, S. “Automatic Classification of Emotion-Related User States in Spontaneous Children’s Speech”, Logos Verlag, Berlin, 2009.