Recognizing Emotion from Singing and Speaking Using Shared Models

Biqiao Zhang, Georg Essl, Emily Mower Provost
Computer Science and Engineering, University of Michigan, Ann Arbor
Email: {didizbq, gessl, emilykmp}@umich.edu

Abstract—**Speech and song are two types of vocal communications that are closely related to each other. While significant progress has been made in both speech and music emotion recognition, few works have concentrated on building a shared emotion recognition model for both speech and song. In this paper, we propose three shared emotion recognition models for speech and song: a simple model, a single-task hierarchical model, and a multi-task hierarchical model. We study the commonalities and differences present in emotion expression across these two communication domains. We compare the performance across different settings, investigate the relationship between evaluator agreement rate and classification accuracy, and analyze the classification performance of individual feature groups. Our results show that the multi-task model classifies emotion more accurately compared to single-task models when the same set of features is used. This suggests that although spoken and sung emotion recognition tasks are different, they are related, and can be considered together. The results demonstrate that utterances with lower agreement rate and emotions with low activation benefit the most from multi-task learning. Visual features appear to be more similar across spoken and sung emotion expression, compared to acoustic features.**

I. Introduction

Speech and song are often considered overlapping forms of vocal expression. For example, in ancient Greek, the words for singing and speaking do not have the distinct meanings they do today [1]. Further, emotion is expressed over both song and speech. Previous works have looked into the acoustic and visual cues in spoken and sung emotional communication [2]–[4], and significant progress has been made in both speech and music emotion recognition [5]–[9]. However, few works have concentrated on building shared emotion recognition models for song and speech. A shared model is desirable because it generalizes across types of vocal communications, which may help combat data scarcity.

In this work, we predict emotion expressed in speech and song using shared models and analyze the commonalities and differences present in emotion expression across these two communication domains. We propose three models for shared emotion recognition from speech and song: (1) a simple model, where a single classifier is built to recognize emotions in both domains; (2) a single-task hierarchical model, where domain classification is performed first, followed by emotion classification of speech and song; (3) a multi-task hierarchical model, where the independent emotion classifiers in (2) are considered together in a multi-task setting. Domain refers to speech and song in this paper. For models (2) and (3), domain information is required only for the training data. Finally, we use two different feature selection techniques, one selecting features on all training data, regardless of domain, and the other selecting features separately for speech and song.

The results show that the multi-task hierarchical model outperforms the other two models when the same set of common features is used, and can either outperform or obtain comparable results compared to models using features specifically selected for each domain. This supports the relatedness of speech and song emotion classification. We find that separate feature selection for song and speech works better than combined feature selection in the single-task hierarchical model, but does not outperform the multi-task hierarchical model. This indicates that emotion expressions in song and speech modulate some features differently, but also share feature modulations. In addition, utterances with lower agreement rate and emotions with low activation benefit more from multi-task learning compared to high agreement and high activation utterances. Classification using individual feature sets demonstrated that visual features are more consistent across speech and song than acoustic features. The novelty of this paper includes: (1) combining speech and song emotion recognition into a unified framework; (2) introducing multi-task learning to emotion recognition from different domains of vocal communications.

II. Related Works

A. Music and Speech Emotion Recognition

Previous works on music and speech emotion recognition have investigated the emotion information encoded in audio [10], [11], video [7], [12] and audio-visual cues [8], [9]. Both acoustic and visual features have been demonstrated useful for speech emotion recognition. Although most works on music emotion recognition depend on acoustic features [6], a study of the use of facial movement to communicate emotion shows the importance of facial expressions in emotional communication during singing performances [13]. In spite of the fast progress of both speech and music emotion recognition, few works have considered emotion recognition from musical (song) and non-musical (speech) vocal communication as related tasks and built shared emotion recognition models.

B. Comparison between Singing and Speaking Emotion Communication

Past researchers have looked at the similarity between music and speech emotion communication. Juslin and Laukka [14] conducted a meta-analysis of studies analyzing speech and music performance. They found that some acoustic features play similar roles in both music and speech emotion communication. For example, anger is associated with fast tempo, high sound level and elevated high-frequency energy in both music and speech. Ilie and Thompson [15], [16] found that the manipulation of certain acoustic features of music and speech, such as loudness and rate, resulted in similar emotion perception. For example, loud excerpts were judged as more pleasant, energetic, and tense. The work of Weninger et al. [17] investigated the shared acoustic features in speech, music and sound and introduced the cross-domain correlation coefficient as a measure of relevance. Their results indicate that there may be shared and divergent properties in emotion communication across domains of expression.

Researchers have analyzed feature-level properties, comparing spoken and sung emotion expressions. Scherer et al. [2] compared emotion expression in singing and speaking. They found that there were significant differences in the acoustics across emotion classes in sung communication. Moreover, they found a high degree of similarity, comparing the patterns of sung expressions and spoken expressions of emotion. Our previous work [18] investigated emotion perception in singing and speaking using within-domain and cross-domain prediction models. The results suggested that activation is perceived more similarly across domains, compared to valence. Furthermore, visual features capture cross-domain emotion more accurately than acoustic features.

Livingstone et al. conducted experiments and analyses on the similarities and differences in the acoustic [3] and visual [4] cues between spoken and sung emotion expression. They found that emotion is conveyed similarly in many acoustic features across speech and song. They also reported differences in several acoustic parameters, including vocal loudness, spectral properties and vocal quality [3]. In [4], they found that emotion-dependent movements of the eyebrows and lip corners are similar between speech and song, yet the jaw movements were coupled to acoustic intensity and thus show difference across spoken and sung emotion expression. They also found that facial expressions conveyed emotion more accurately than vocal expressions for sung stimuli. These findings provide evidence for the links between speech and song, and give support to the notion of building a shared model for speech and song emotion recognition. However, it is not yet clear how the similarities and differences between spoken and sung expressions of emotion can be used to build a shared model.

C. Multi-class and Multi-task SVM

Support Vector Machine classifiers (SVM) were originally designed for binary classification. There are currently two types of approaches for multi-class SVM: one builds and combines multiple binary classifiers, while the other considers all data in one optimization problem [19], [20]. Typical methods of the former approach includes one-against-all [21], one-against-one [22], [23] and directed acyclic graph SVM (DAGSVM) [24]. Experiments comparing these two approaches showed that the combinations of binary classifiers outperformed the all-in-one multi-class optimization in both classification accuracy and training/testing time, and demonstrated that one-against-one method and DAG are more suitable for practical use [25].

In [26], Evgeniou et al. extended single-task SVMs to multi-task scenario. This extension is based on the assumption that there are T related tasks that share the same space. It solves multiple tasks together by imposing a regularization constraint on the average model and controlling the differences between the tasks using additional model parameters. Empirical results show that this multi-task SVM outperforms earlier multi-task learning methods as well as single-task SVM. Evgeniou, Micchelli and Pontil [27], [28] improved the idea of [26] by proposing multi-task kernels. They developed multi-task kernels to extend the single-task learning methods that use kernels to multi-task learning. These methods all require that the tasks are label-compatible (tasks sharing the same set of labels). The work of [29] lessened the requirement for a shared space between tasks. Ji and Sun proposed a multi-task multi-class SVM based on the single-task version of multi-class SVM that considers all data in one optimization formulation [20], [25]. The main advantage of this method is that it can support label-incompatible multi-task learning (tasks having different sets of labels), since the multi-class classification problem does not depend on a set of binary classifiers.

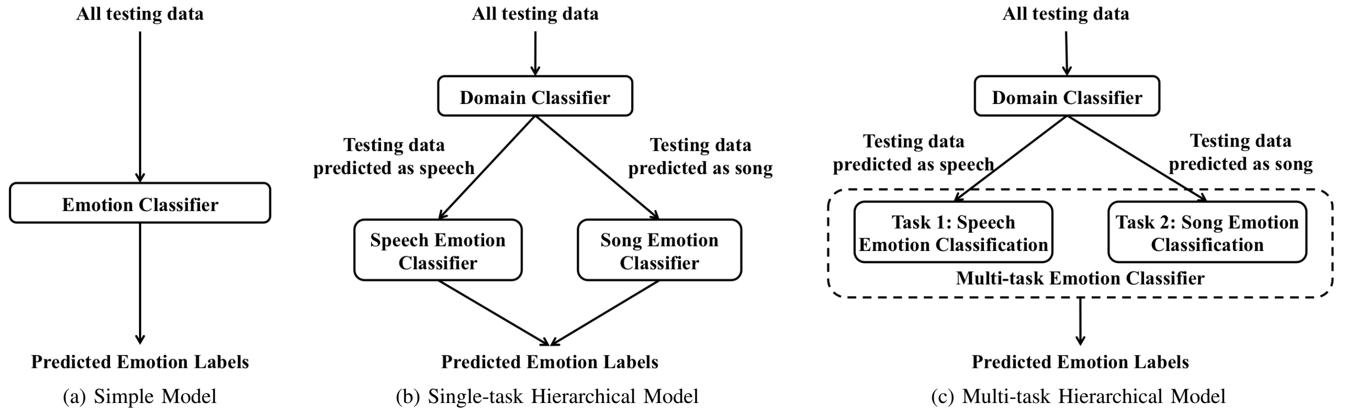
III. Datasets

We use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [30] and the University of Michigan Song and Speech Emotion Dataset (UMSSED) [18].

A. The RAVDESS Dataset

The RAVDESS dataset is used in [3], [4] for investigating the similarities and differences in acoustic and visual signals of emotional singing and speaking. It contains audio-visual recordings of 24 performers (12 male, 12 female) speaking and singing the same two sentences with different emotions at normal and strong emotional intensity, each with two repetitions. The speech recordings consist of 8 emotions, including neutral, calm, happy, sad, angry, fearful, disgust and surprise. The song recordings consist of the first 6 emotions. Three melodies were composed for the singing performances: one each for neutral, positively valenced and negatively valenced emotions. The melodies only differ in two notes in the middle.

The emotion content of the dataset was evaluated by 350 human participants, each utterance receiving 10 ratings from 10 different participants. The participants were asked to identify the emotion expressed by the performer from the set of the target emotions, or indicate none is correct. An agreement rate ranging from 0 to 1 was calculated for each utterance.


All testing data
Emotion Classifier
Predicted Emotion Labels
All testing data
Domain Classifier
Testing data predicted as speech
Testing data predicted as song
Speech Emotion Classifier
Song Emotion Classifier
Predicted Emotion Labels
All testing data
Domain Classifier
Testing data predicted as speech
Testing data predicted as song
Task 1: Speech Emotion Classification
Task 2: Song Emotion Classification
Multi-task Emotion Classifier
Predicted Emotion Labels

(a) Simple Model (b) Single-task Hierarchical Model (c) Multi-task Hierarchical Model

Fig. 1. Three emotion classification models: (a) simple model, (b) single-task hierarchical model and (c) multi-task hierarchical model. In (a), testing data enter a single emotion classifier trained on all training data, regardless of domain. In (b), a domain classifier and two separate domain-specific emotion classifiers are trained. Testing data is first classified into speech or song domain, and then enter the corresponding classifier accordingly. In (c), a domain classifier and a multi-task classifier are trained. The task which the testing data belong to is decided by the domain classification result.

A score of 1 means total consensus between evaluators and the target emotion, while a score of 0 means complete lack of consensus. The evaluators also rated the emotional intensity and the emotional genuineness of the performer.

We only used utterances with normal intensity to avoid using exaggerated performances. We further decrease the dataset to address two concerns: (1) there are two emotions unique to speaking and (2) one performer has only speaking data. These characteristics introduce dependencies between domain, emotion, and performer that could bias the results. Therefore, we eliminated the disgust and surprise utterances and dropped the performer with only speaking data. This results in 1104 audio-visual utterances in total (2 domains × 23 performers × 6 emotions × 2 sentences × 2 repetitions). The target emotion of performers is used as ground truth because the perceived emotion labels are not yet available. The average agreement rate of the target emotion is 0.69. The chance rate of agreement and classification accuracy are both 16.7%. See [3], [4], [30] for additional details.

B. The UMSSED Dataset

The UMSSED Dataset consists of audio-visual recordings of three performers (1 female, 2 male) singing and speaking seven semantically neutral sentences [18]. Each sentence is embedded into passages associated with angry, happy, neutral and sad emotions to facilitate emotionally evocative performances. Seven melodies were composed to match the passages, one each for the seven sentences. The four emotion variations of each target sentence are accompanied by the exact same melody. This results in 168 utterances in total (2 domains × 3 performers × 4 emotions × 7 sentences). The target sentence was segmented out from the remainder of the passage and evaluated using Amazon Mechanical Turk by 183 evaluators. Each utterance was evaluated by 20.9 ± 1.7 participants. The evaluators were required to assess the primary emotion of the utterance from the set of angry, happy, neutral, sad and other, and to rate the valence, activation and dominance level. We used the target emotion of performers as ground truth to match the RAVDESS data. The average agreement rate of the target emotion is 0.62. The chance rate of agreement and classification accuracy are both 25%. Please see [18] for additional database details.

IV. Methodology

A. Feature Extraction

Audio-visual features were extracted to predict emotion expressed in song and speech.

1) Acoustic Features: We used the low level descriptors (LLDs) described in the Interspeech 2013 Computational Paralinguistics Evaluation (ComParE) feature set [31]. The feature set includes 4 energy, 41 spectral, 14 cepstral (MFCC) and 6 voicing-related LLDs. We used openSMILE [32] to extract the 65 LLDs. We applied statistics including mean, standard deviation, max, min, range, interquartile range, mean absolute deviation, skewness and kurtosis to the non-silence part of the LLDs and delta LLDs to generate utterance-level acoustic features. This resulted in 1170 features.

2) Visual Features: We extracted the probabilities of 20 facial action units (AUs), left and right unilateral AUs of 3 AUs within the 20 (Lip Raise, Lip Corner Pull and Dimpler), and 2 AU groups (Fear Brow and Distress Brow) at the frame-level, using CERT [33]. Action units are the fundamental actions of individual muscles or groups of muscles. We applied the same set of statistics as above to the frame-level LLDs and delta LLDs, which results in 504 utterance-level features.

B. Classification Models

We present three emotion classification models as shown in Fig. 1. The simple model (Fig. 1a) creates a single classifier, independent of domain. The two hierarchical models (Fig. 1b and 1c) use domain during training. The single-task model (Fig. 1b) trains a separate emotion classifier for each domain. The multi-task model (Fig. 1c) trains a multi-task classifier to jointly predict emotion across both domains. In the testing phase, the testing data are separated based on the predicted domain. The data are analyzed using the classifier corresponding to the estimated domain (Fig. 1b and 1c).

We adopted the directed acyclic graph SVM (DAGSVM) [24] as our single-task multi-class emotion classifier. DAGSVM is identical to the one-against-one SVM in the training phase. It constructs a binary classifier for each pair of classes, thus for a multi-class problem with k classes, $k(k-1)/2$ classifiers are trained. In the testing phase, the DAGSVM uses a rooted binary directed acyclic graph that contains $k(k-1)/2$ internal nodes and k leaf nodes. Each internal node represents a binary classifier. Each test case starts from the root node and moves to the left or right depending on the output of the classifier until a leaf node indicating the predicted class is reached. One advantage of DAGSVM is that it has a shorter testing time than the one-against-one method, but still has similar classification performance [25]. We used the implementation of [25], which is built on LIBSVM [34].

We used the regularized multi-task SVM [26] for the second stage of our multi-task hierarchical model. We did not select the more recent multi-task SVM in [29] because its main advantage is the ability to support label-incompatible tasks, while in this paper, the tasks are label-compatible.

The method of Evgeniou et al. [26] extends single task SVMs to multi-task scenario by explicitly taking the relationship between learned weights in different tasks into account. For T related tasks that share the same feature space X and label space Y, this method learns T classifiers $w_1, ..., w_T$, where w_t is specific to task t. The w_t can be written as $w_0 + v_t$, where w_0 is the shared information across all tasks. The values of w_0 and v_t can be obtained by solving the following optimization function

$$\min_{w_0, v_t, \epsilon_{t,i}} \frac{\lambda_1}{T}||v_t||_2^2 + \lambda_2||w_0||_2^2 + \sum_{t=1}^{T}\sum_{i=1}^{m_t}\epsilon_{t,i}, \tag{1}$$

$$s.t. \forall t, i, y_{t,i}(w_0 + v_t) \cdot x_{t,i} \geq 1 - \epsilon_{t,i}, \epsilon_{t,i} \geq 0. \tag{2}$$

Here, t is the index of the task, i is the index of the utterance, $\epsilon_{t,i}$ is the error term, λ_1 and λ_2 are two non-negative constants that control the relationship between the tasks. Solving the dual form of the above optimization function is equivalent with the standard C-Support Vector Classification (C-SVC), but replacing the kernel function with

$$K_{\phi(i)\phi(j)}(x_i, x_j) = (\delta_{\phi(i)\phi(j)} + \frac{1}{\mu})k^*(x_i, x_j), \tag{3}$$

where $k^*(x_i, x_j)$ stands for regular kernel functions, $\phi(i)$ is the task x_i belongs to, $\delta_{\phi(i)\phi(i)} = 1$ if $\phi(i) = \phi(j)$ and 0 otherwise, and μ is a parameter for controlling the similarities between tasks. In the two extreme cases, the tasks are the same when $\mu \to 0$, and are completely decoupled if $\mu \to \infty$.

There are two parameters that must be selected: the cost parameter C for the training error, as in the standard C-SVC, and the parameter μ. In this work, we used the RBF Gaussian kernel for $k^*(x_i, x_j)$, and thus introduced another parameter γ that controls the bandwidth of the Gaussian function.

V. Experimental Setup

A. Performer Normalization

For the 1674 utterance-level features, we used performer dependent z-normalization across all utterances of the same performer such that each feature of each individual performer has zero mean and standard deviation of one. This method is commonly adopted in emotion recognition to mitigate the intrinsic differences in the vocal characteristics and facial muscle movements of speakers [9], [35]

B. Cross Validation

We used different cross validation paradigms for the two datasets to measure the performance of the models. For UMSSED, we used leave-one-performer-out cross validation (one performer as testing data, others used for training data). As the RAVDESS has very limited lexical variability, we used leave-one-performer-and-sentence-out cross validation to avoid overfitting to specific lexical content. In each round, one sentence from one performer serves as testing data. The training data are composed of the other performers and other sentence. Each performer is associated with two cross-validation folds, one for each sentence. The parameter tuning process was performer-independent and based only on the training set. The parameters of the models were selected using a grid search by optimizing the 5-fold cross-validation accuracy of the training data.

C. Feature Selection

We reduced the dimensionality of the feature set by selecting task-related features using Information Gain [36] on the training data. The number of features selected was decided by maximizing the average cross-validation accuracy of the training data. We selected features across domain (song and speech) for the simple model and the emotion classification phase of both the single-task hierarchical model (st-hier) and the multitask hierarchical model (mt-hier). We selected domain-specific features for the single-task hierarchical model with separate feature selection (st-hier-sfs). The mean and standard deviation of the number of features selected for domain and emotion classification are 150 ± 141 and 283 ± 144, respectively.

VI. Results and Analysis

A. Performance of Different Models

Table I shows the emotion classification accuracy of the simple model, single-task hierarchical model (st-hier), single-task hierarchical model with separate feature selection (st-hier-sfs) and the multi-task hierarchical model (mt-hier) on the RAVDESS and UMSSED datasets. The domain classification used for the latter two models has an accuracy of 99% on RAVDESS and 97% on UMSSED. We can observe that when all three models use the same features, the mt-hier model outperforms the other two in both datasets. The difference is statistically significant for RAVDESS, where significance is asserted at $\alpha = 0.05$ (paired t-test, p=0.002 for simple and

TABLE I
EMOTION CLASSIFICATION ACCURACY OF DIFFERENT MODELS (%)

Dataset	simple	st-hier	st-hier-sfs	mt-hier
RAVDESS	81.61	81.25	83.15	83.15
UMSSED	70.83	70.83	71.43	74.40

mt-hier, p=0.044 for st-hier and mt-hier). The performance increase for mt-hier is not significant for UMSSED. This may be due to the fact that this dataset only contains three performers. This result provides evidence that song and speech emotion classification are different, but related, tasks. By adding the relationship into the model, we can outperform models that either consider them as the same task, or consider them as completely unrelated tasks.

With separate feature selection for song and speech emotion classification, the st-hier-sfs model works better than st-hier in both datasets (significant for RAVDESS, paired t-test, p=0.046). However, the mt-hier model using features selected on combined song and speech data can achieve the same accuracy as the st-hier-sfs in the RAVDESS dataset, and outperforms all other methods in the UMSSED dataset. This may indicate that the features that are most important to emotion classification in song and speech have some differences, but also share similarities.

B. Performance and Agreement Rate

We are interested in the relationship between classification accuracy and performer-evaluator agreement rate, and the benefit multi-task learning brings for utterances with different agreement rates. We compared the agreement rates between correctly classified and incorrectly classified utterances using the simple model and mt-hier model in both RAVDESS and UMSSED datasets. We found that correctly classified utterances have significantly higher agreement rates than incorrectly classified utterances using two-sample t-test under significance level of 0.01, regardless of model and dataset. Table II shows the emotion classification accuracy of low (0 to one third, inclusive, noted as [0,1/3]), medium (one third to two third, open and inclusive, noted as (1/3,2/3]) and high agreement rate ((2/3,1]) utterances using simple model and mt-hier model in both RAVDESS and UMSSED datasets. It can be observed that classification accuracy and agreement rate are positively correlated. Comparing the accuracy of simple model and mt-hier model, we found that although all low, medium, and high agreement utterances have higher classification accuracy in mt-hier, low and medium utterances benefit more from using multi-task approach.

Fig. 2 visualizes the confusion matrix of the classification results on RAVDESS dataset (2a) and UMSSED (2b) dataset of all utterances from both domains, using the simple model and mt-hier model. It can be seen that emotions with high activation (e.g. angry, happy, fearful) are easier to classify than emotions with low activation (e.g. neutral, sad) for the simple model. One interesting observation is that emotions with low activation benefit more from the multi-task approach than emotions with high activation. For example, the accuracy of the neutral class is boosted to 81% from 62% for speech utterances, and the accuracy of the sad class is increased by 10% for the sung utterances in the UMSSED dataset when multi-task approach is used, as shown in Fig. 2b.

TABLE II
EMOTION CLASSIFICATION ACCURACY OF LOW, MEDIUM AND HIGH AGREEMENT RATE UTTERANCES (%)

	RAVDESS		UMSSED	
	simple	mt-hier	simple	mt-hier
Low agreement [0,1/3]	70	74.17	60.61	63.64
Medium agreement (1/3,2/3]	77.5	80.39	65.96	70.21
High agreement (2/3,1]	85.55	85.84	77.27	80.68

C. Performance of Individual Feature Groups

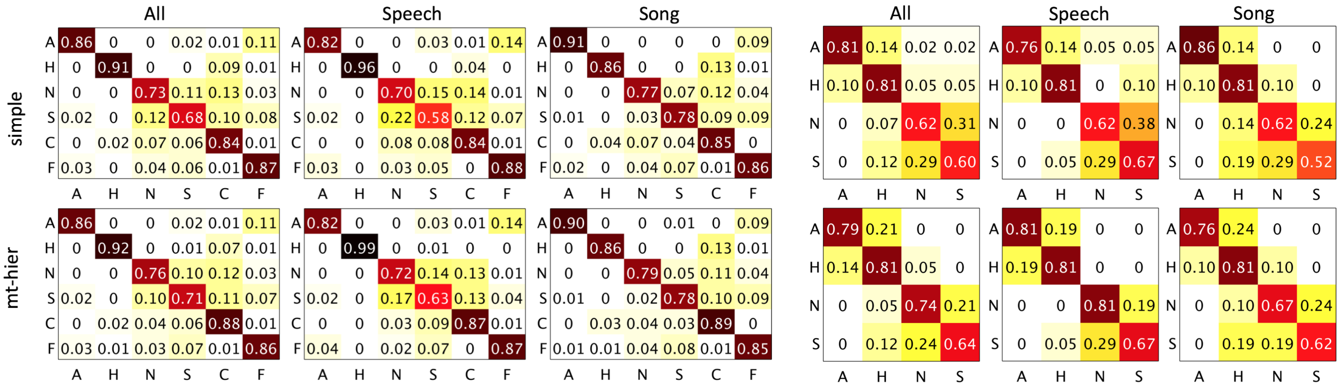
We group features by the type of their low level descriptor (LLD) and show the domain classification and emotion classification accuracy in Table III and Table IV, respectively. Acoustic features are grouped into energy-related, spectral, MFCC and voicing-related features based on the categorization in [31]. We separate out RASTA-style auditory spectrum from spectral features since they are calculated differently. Previous research has demonstrated that there are differences between emotion associated with the upper and lower face [37]. Therefore, we categorized facial action unit features into upper-face and lower-face action units. We compare features for emotion classification using single-task SVM classifiers were trained using all utterances, only spoken utterances and only sung utterances. We did not use hierarchical models to eliminate the influence of domain classification accuracy on emotion classification accuracy.

As shown in Table III, all five groups of acoustic features can predict domain accurately. Among them, voicing related features have the highest accuracy across both datasets. On the other hand, the prediction accuracies of visual features are relatively low. Lower face action units are better distinguishers of domain, compared with upper face action units.

In Table IV, we notice that in both datasets, energy-related features, spectral features and voicing features can predict emotion in song more accurately than in speech, which may suggest that they contain more emotion-specific patterns related to the singing voice. In addition, the lower face visual features are effective across both domains, outperforming the models trained on only speech or only song. The upper face features exhibit this pattern only for the RAVDESS data. This suggests that the visual features may exhibit more similarities across domain, compared to the audio features. Other patterns of accuracy are not very consistent between the two datasets. Possible explanations include that the number of emotion labels of the two datasets are not the same, and that the melody matching methods are also different.

VII. CONCLUSION AND DISCUSSION

In this paper, we proposed three shared emotion recognition models for speech and song: the simple model, the single-task hierarchical model and the multi-task hierarchical model. We studied the commonalities and differences present in



(a) Confusion Matrix: RAVDESS

(b) Confusion Matrix: UMSSED

Fig. 2. Confusion matrix of the classification results on (a) RAVDESS and (b) UMSSED of all utterances, utterances from speech and utterances from song using simple model and mt-hier model. The darker the color, the higher the accuracy/confusion. A: Angry, H: Happy, N: Neutral, S: Sad, C: Calm, F: Fearful.

TABLE III
DOMAIN CLASSIFICATION ACCURACY OF EACH FEATURE GROUP (%)

Feature group	RAVDESS	UMSSED
Energy	94.75	95.83
Spectral	98.19	90.48
MFCC	97.46	95.83
Voicing	99.09	98.21
Rasta	96.69	89.29
Upper face	71.56	58.93
Lower face	78.61	80.95

TABLE IV
EMOTION CLASSIFICATION ACCURACY OF EACH FEATURE GROUP (%). PERFORMANCES GIVEN BY SINGLE-TASK SVM CLASSIFIERS TRAINED USING ALL UTTERANCES (ALL), ONLY SPOKEN UTTERANCES (SPEECH) AND ONLY SUNG UTTERANCES (SONG).

	RAVDESS			UMSSED		
Feature group	All	Speech	Song	All	Speech	Song
Energy	37.86	34.78	40.22	60.12	55.95	61.90
Spectral	50.82	48.01	51.99	60.12	52.38	61.90
MFCC	53.80	48.73	62.68	47.02	51.19	46.43
Voicing	45.38	38.41	53.62	43.45	39.29	51.19
Rasta	34.33	30.43	32.97	51.19	61.90	44.05
Upper face	71.83	69.20	70.83	55.95	48.81	58.33
Lower face	66.85	65.40	64.67	64.88	52.38	53.57

emotion expression across these two communication domains by comparing the performance of these difference settings, investigating the relationship between agreement rate and classification accuracy, and analyzing the classification performance of individual feature groups.

The fact that classification accuracy benefits from multi-task learning suggests that despite the differences between speech and song emotion recognition, they are related and can be considered using a shared model. The single-task hierarchical model performs better when using features selected separately for speech and song compared to combined feature selection, yet it does not exceed the multi-task hierarchical model. This may indicate that although there are some differences in the important features for speech and song emotion recognition, more similarities can be found. Classification accuracy of utterances with lower agreement rate increases more with multi-task learning. Therefore, it is possible that multi-task learning can help distinguish emotions in ambiguous situations.

Classification using individual feature sets demonstrated that acoustic features are better distinguishers of domain than visual features. The emotion classifier using visual features from both domains is more accurate than the one operating on only spoken or only sung data (lower face features for RAVDESS and UMSSED and upper face features for RAVDESS). This suggests that visual features are more similar across domain, compared to acoustic features. This is in accordance with the findings in [18] that visual features work better in cross-domain emotion prediction. Among visual features, domain classification using lower face action units performs better than classification using upper face action units, which may suggest that upper face action units are more similar in speech and song. This also agrees with the findings in [4] that emotion-dependent movements of the eyebrows and lip corners are similar between speech and song, yet the jaw movements show difference across speech and song emotion expression.

A limitation of this paper is that we only adopted datasets with acted emotion. We plan to use more datasets, including non-acted datasets, and perform cross-corpus analysis in our future work for better generalizability. In this paper, we discussed the situation in which the multi-task model uses the same features as the simple model, where features are selected on all the training data, regardless of domain. There have been methods for multi-task feature learning as described in [38]–[40]. Therefore, it would be interesting to learn multi-task features for speech and song emotion recognition and compare them with features selected using traditional single-task feature selection methods.

ACKNOWLEDGEMENT

We would like to express our appreciation to Dr. Steven Livingstone, who shared the RAVDESS dataset with us before its official release.

References

[1] L. Stamou, "Plato and aristotle on music and music education: Lessons from ancient greece," *International Journal of Music Education*, no. 1, pp. 3–16, 2002.

[2] K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salomão, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Computer Speech & Language*, 2013.

[3] S. R. Livingstone, K. Peck, and F. A. Russo, "Acoustic differences in the speaking and singing voice," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1. Acoustical Society of America, 2013.

[4] S. R. Livingstone, W. F. Thompson, M. M. Wanderley, and C. Palmer, "Common cues to emotion in the dynamic facial expressions of speech and song," *The Quarterly Journal of Experimental Psychology*, pp. 1–19, 2014.

[5] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *International Society for Music Information Retrieval*. Citeseer, 2010, pp. 255–266.

[6] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, p. 40, 2012.

[7] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *International Conference on Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 2474–2477.

[8] N. Sebe, I. Cohen, T. Gevers, and T. S. Huang, "Emotion recognition based on joint visual and audio cues," in *International Conference on Pattern Recognition*, vol. 1. IEEE, 2006, pp. 1136–1139.

[9] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1057–1070, 2011.

[10] C. Baume, "Evaluation of acoustic features for music emotion recognition," in *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[11] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[12] Y. Kim and E. Mower Provost, "Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition," in *ACM International Conference on Multimedia*, 2014.

[13] L. R. Quinto, W. F. Thompson, C. Kroos, and C. Palmer, "Singing emotionally: A study of pre-production, production, and post-production facial expressions," *Frontiers in Psychology*, vol. 5, 2014.

[14] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychological Bulletin*, vol. 129, no. 5, p. 770, 2003.

[15] G. Ilie and W. F. Thompson, "A comparison of acoustic cues in music and speech for three dimensions of affect," *Music Perception: An Interdisciplinary Journal*, vol. 23, no. 4, 2006.

[16] ——, "Experiential and cognitive changes following seven minutes exposure to music and speech," *Music Perception: An Interdisciplinary Journal*, vol. 28, no. 3, 2011.

[17] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: What speech, music, and sound have in common," *Frontiers in Psychology*, vol. 4, 2013.

[18] B. Zhang, E. Mower Provost, R. Swedberg, and G. Essl, "Predicting emotion perception across domains: A study of singing and speaking," in *Proceedings of AAAI*, 2015.

[19] J. Weston and C. Watkins, "Multi-class support vector machines," Citeseer, Tech. Rep., 1998.

[20] K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," *Machine Learning*, vol. 47, no. 2-3, pp. 201–233, 2002.

[21] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel, Y. LeCun, U. A. Muller, E. Sackinger, P. Simard *et al.*, "Comparison of classifier methods: a case study in handwritten digit recognition," in *International Conference on Pattern Recognition*. IEEE Computer Society Press, 1994, pp. 77–77.

[22] J. Friedman, "Another approach to polychotomous classification," Technical report, Department of Statistics, Stanford University, Tech. Rep., 1996.

[23] U. H.-G. Kreßel, "Pairwise classification and support vector machines," in *Advances in kernel methods*. MIT Press, 1999, pp. 255–268.

[24] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification." in *nips*, vol. 12, 1999, pp. 547–553.

[25] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[26] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 109–117.

[27] C. A. Micchelli and M. Pontil, "Kernels for multi-task learning," in *Advances in Neural Information Processing Systems*, 2005, pp. 921–928.

[28] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," in *Journal of Machine Learning Research*, 2005, pp. 615–637.

[29] Y. Ji and S. Sun, "Multitask multiclass support vector machines," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 512–518.

[30] S. R. Livingstone, K. Peck, and F. A. Russo, "Ravdess: The ryerson audio-visual database of emotional speech and song," in *Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science*, 2012.

[31] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013.

[32] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[33] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 298–305.

[34] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[35] C. Busso and S. S. Narayanan, "The expression and perception of emotions: comparing assessments of self versus others." in *INTERSPEECH*, 2008, pp. 257–260.

[36] T. Cover and J. Thomas, "Elements of information theory," 1991.

[37] E. D. Ross, C. I. Prodan, and M. Monnot, "Human facial expressions are organized functionally across the upper-lower facial axis," *The Neuroscientist*, vol. 13, no. 5, pp. 433–446, 2007.

[38] A. Evgeniou and M. Pontil, "Multi-task feature learning," *Advances in neural information processing systems*, vol. 19, p. 41, 2007.

[39] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[40] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient $l_{2,1}$-norm minimization," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 339–348.