JOINT GENDER-, TONE-, VOWEL- CLASSIFICATION VIA NOVEL HIERARCHICAL CLASSIFICATION FOR ANNOTATION OF MONOSYLLABIC MANDARIN WORD TOKENS

Saurabh Garg[a,d], *Ghassan Hamarneh*[b], *Allard Jongman*[c], *Joan Sereno*[c], *and Yue Wang*[d]

[a]Pacific Parkinson's Research Centre, University of British Columbia, Canada

[b]Medical Image Analysis Lab, Simon Fraser University (SFU), Canada

[c]KU Phonetics and Psycholinguistics Lab, Department of Linguistics, University of Kansas (KU), USA

[d]Language and Brain Lab, Department of Linguistics, Simon Fraser University, Canada

ABSTRACT

The automatic annotation of Mandarin monosyllabic audio word tokens remains an important yet challenging issue in phonetics research. In this work, we address this annotation task via a novel subcategories-classification framework that not only performs word identification via the joint classifications of vowel and tone subcategories, but also performs gender discrimination of the speaker, which stands in contrast to previously proposed methods for Mandarin speech that focused only on tone-, vowel-, or gender- classification. We also propose a novel hierarchical classification algorithm to boost overall classification performance. Extensive experimental results show that our approach yielded superior performance in both cases of adequate and very limited training data. When trained using data from only one female and one male speaker, our approach also yielded the best classification accuracy in all subcategories of the token annotation problem, achieving an F1-score of 0.742 as opposed to 0.705 as achieved by the second competing approach.

Index Terms— monosyllabic Mandarin tokens, phonetics, hierarchical classification, label fusion, multi-class SVMs

1. INTRODUCTION

The automatic annotation (AA) of monosyllabic word tokens extracted from audio recordings is a fundamental yet critical step even in the current modern era due to the needs cast by phonetic studies [1, 2], clinical therapy designs, and treatment evaluations for speech and hearing impairment [3].

In this work, we aim to annotate segmented audio tokens of Mandarin words for a series of prospective Mandarin tone studies being conducted at SFU and KU [4, 5]. In particular, given an extracted audio token of a Mandarin word articulated by a speaker, our goal is to assign to this token a set of annotations, namely: vowel, tone, and gender. These annotations are then used to facilitate subsequent intelligibility tests and group analyses of the differences in the acoustic features of individual words for the aforementioned prospective studies.

Our goal is challenging to achieve for several reasons. Firstly, unlike the common continuous-speech-based automatic speech recognition, the annotation of monosyllabic words is challenging by the lack of contextual information. In particular, the discrimination of certain words in isolation is difficult even for human raters based on our rater perception data (see Section 2), mostly due to the challenge of discriminating the Mandarin rising tone from the dipping tone. Secondly, inter-speaker variability observed in our dataset is high because of the two speaking styles involved in this work such that one single word could be produced in a hyper-articulated, clear speech manner that is different from a conversational, plain speaking style. Thirdly, the number of word categories and the total dataset is generally limited, as compared to data from continuous speech studies. Hence, when limited training data is available (which is the case for small-scale analyses of hand-selected monosyllabic Mandarin speech tokens), and when pretrained models are not available for transfer learning, methods based on deep-learning (e.g. those based on convolutional neural networks (CNN) [6, 7]) may not be suitable, as our experimental results suggest.

To circumvent these challenges, we propose a new formulation of AA as the joint classifications of three subcategories, namely: i) identification of each token as one of 3 vowel classes, ii) identification of each token as one of 4 tone classes; and iii) identification of the speaker's gender. To the best of our knowledge, this formulation has not been proposed in the literature before (Table 1). By framing annotation as the labelings of 3 subcategories, we can engineer our framework by following the Mandarin phonetic model as closely as possible. Additionally, we propose a novel hierarchical classification algorithm using support vector machine (SVM) that would allow us to employ gender-specific vowel classifications and vowel-specific tone classifications. By formulating the classification subproblems hierarchically, we leverage domain-specific knowledge to custom-design feature-classifier combinations to further improve classification performance, as our experimental results show.

2. MATERIALS

For the purpose of answering specific research objectives set in [4, 5], the speech tokens were carefully chosen to form a dictionary of 12 Mandarin words, each of which belongs to one of 3 vowel classes (/ɜ/, /i/ and /u/), and belongs to one of 4 tones classes (Tone 1: level; Tone 2: rising; Tone 3: dipping; Tone 4: falling). Furthermore, the production of each token

Table 1: Relation to prior work for AA of Mandarin monosyllabic speech tokens.
'V', 'T', and 'G' respectively denotes whether each paper addressed classification(s) of vowel, tone, and gender; S, T, k respectively denotes the size of the training set, size of the test set, number of classes handled; M and F denote male and female.

Ref.	V	T	G	Classifier employed	S	T	k
[8]	✗	✓	✗	Logistic regression, SVM with RBF, NN	7549	300	6 tones
[9]	✓	✗	✗	SVM with RBF, DNN	18 sentences (1M, 1F)	NA	2
[10]	✗	✓	✗	3-layer neural network (NN)	6000 (8M, 8F)	750	4 tones
[11]	✗	✓	✗	Deep neural network (DNN)	7549	300	5 tones
[7]	✗	✓	✗	CNN	50 utterances from 1M, 1F	1 utterance from 40 subjects	≈40
[12]	✗	✓	✗	Multi-feed-forward NN	1539	670	4 tones
[13]	✗	✓	✗	SVM with radial basis function (RBF) kernel	10080	10080	4 tones
[14]	✓	✗	✗	SVM with linear, 3rd order Polynomial and RBF	1176	391	5 vowels
[15]	✗	✗	✓	SVM (choice of kernel not mentioned)	70	240	6+3 age groups
[16]	✗	✗	✓	Logistic Regression, Random Forest, AdaBoost	3616	452	2 gender

Table 2: Feature sets we employed for each subcategory.

Feature id	**Description**
	Gender [14]
G1-G2	Absolute and relative jitter
G3-G6	Mean, standard deviation, mode and median of Fundamental Frequency (F0)
G7-G8	Minimum and maximum value of pitch contour over time
G9	Inter-quartile range of F0
G10-G12	Coefficients of second order fitted polynomial on the estimated pitch contour (p)
G13-G14	Positions of the minima and maxima on p
G15	Ratio of F0 to F1
	Tones [17]
T1-T3	Coefficients of second order polynomial function fitted on the estimated pitch contour
T4-T5	Relative positions of minima and maxima on p
T6-T10	Slopes of p
T11-T14	$(a_{max}$-$b_{min})$, $(c_{max}$-$b_{min})$, $(c_{max}$-$a_{min})$, $(a_{max}$-$c_{max})$, where a and c is respectively the first and fourth quartile; b is the union of the second and third quartiles; and the min-subscript (max-subscript) denotes the minimum (maximum) value in that quartile.
	Vowels [14]
V1-V16	Mean of Mel-frequency cepstrum coefficients (MFCCs) over different frames
V17-V19	Median of the F0, F1, F2 frequencies over time

was recorded in isolation (as opposed to continuous speech), and articulated in two speech styles: conversational and clear.

The recruited n=21 speakers consist of 9 males and 12 females who were born and raised in Northern China or Taiwan at least during the first 18 years of their lives. During recording, each speaker articulated each word token at least 10 times in a random order, each articulated in one speech style. In total, 2948 utterances were recorded and were manually annotated by two other native Mandarin speakers. As noted in Section 1, the perception of several tokens with similar tones (e.g. /ɜ2/ vs. /ɜ3/; /u2/ vs. /u3/) is challenging. As our experiment on inter-rater variability shows, the disagreement in the 12-word classification sub-problem can be as high as 17%. Therefore, to ensure the quality of the labels, tokens whose annotation labels disagree were excluded.

3. METHODS

3.1 Feature extraction of speech tokens

Pre-processing: Following standard procedure [18], the mono-channel data is first resampled to 16,000 Hz. Next, the voiced component from the audio signal is separated from background noises by applying a threshold operation on the short-time energy curve [18] of the audio signal. The threshold value was empirically set as 0.05.

Estimating pitch contours: To extract pitch-contours reliably, we integrated two standard approaches (auto-correlation and cepstrum-based) as follows. First, each of the pitch contours generated from these two approaches is median-filtered to remove sudden changes in the pitch estimation over time. Then, the two smoothed contours are analyzed frame-by-frame to obtain the final pitch value $p(i)$. We do so by measuring the similarity of the pitch values at each time frame between the two contours. When the measured similarity is high, we compute $p(i)$ as the average of the two pitch values. When the similarity is below an empirically tuned threshold, $p(i)$ is set as $p(i\text{-}1)$, i.e. the value estimated from the previous frame is used. Lastly, if a pitch value of either approaches is missing as it cannot be estimated due to the common "creakiness" problem [19], linear interpolation is employed whereby $p(i)$ is computed from the two nearest temporal frames where the pitch values are available.

Feature extraction: based on the literature, we extracted a standard set of features for each of the subcategories as listed in Table 2. For MFCC-based features, quantile-based cepstral normalization [20] was used. Lastly, all features are normalized so that they have zero mean and unit variance.

3.2 Hierarchical subcategory-classifiers

We propose a novel hierarchical classification algorithm to tackle our 3-subcategory classification in a joint, integrated framework. Our pipeline (Fig. 1) is motivated by the observation that the defined features of tones and vowels are interdependent. In particular, the length and/or shape of a tone contour may vary depending on its vowel context. For instance, the length of the contour for vowel /ɜ/ tends to be shorter than that for /u/, presumably due to the addition of an initial glide /u/. We thus employ an ensemble of hierarchical classifiers to model these inter-dependencies.

Our proposed classification algorithm is shown in Fig. 1. At the top level, three **generic** classifiers (one for each of the vowel, tone and gender subcategories) are trained using the input training token samples. For each of these classifiers, features specifically developed for each subcategory are extracted from each word token according to Table 2.

In addition to these three generic classifiers, we also construct a second layer of hierarchical classifiers that consist of three more sets of subcategory-specific classifiers: 1) gender-specific vowel classifiers, 2) gender-specific tone classifiers, and 3) vowel-specific tone classifiers.

In training these subcategory-specific classifiers, the training audio tokens are then split into different subgroups. Specifically, to train a gender-specific vowel classifier, the training tokens are split by gender, and a vowel classifier is then trained on the male speech tokens only to form a male-specific vowel classifier, and similarly, another vowel classifier is trained on the female tokens to yield a female-specific vowel classifier. Under this strategy, the variations in the vowel features due to gender differences can be minimized and these second-level classifiers will be trained to learn gender-specific class boundaries that may better differentiate between vowel classes within each gender than those learned by the generic vowel classifier. Likewise, we also build another set of tone classifiers for each of the gender groups. Lastly, to model inter-dependencies between tone-vowel combinations, we also construct a third level of vowel-specific tone classifiers.

Once training is done, testing is done in a cascaded manner where class predictions from all classifiers of all levels are computed. However, predictions from the lower levels are chosen based on the top level. In particular, given a test audio token to be annotated, the gender classifier at the first level is used to predict its gender class. If the predicted gender is male, then only the class predictions computed by the male-specific classifiers are employed to make a prediction for its tone and vowel subcategories.

When all levels of relevant classifiers have made their predictions, three sets of subcategory-specific predictions are available. More specifically, for tone prediction, intermediate predictions are obtained from the gender-specific tone and vowel-specific tone classifiers. In the ideal case where the test token falls in the same distribution as the training tokens that were used to train a generic tone classifier, the prediction of this generic tone classifier can be entirely trusted. However, since it is unlikely that a highly discriminate feature set and a single classifier can produce a highly accurate prediction, we boost its performance by fusing its labels with those acquired from the subcategory-specific classifiers. The final annotation label for the tone subcategory is therefore decided by a majority vote on the predictions made by these three independent classifiers, i.e. the generic tone classifier, gender-specific tone classifier, and vowel-specific tone classifier. Similarly, we also have two different label predictions from the two levels of vowel classifiers, i.e. the generic vowel classifier and gender-specific vowel classifier. The final vowel label of the test token is thus decided based on a weighted majority voting of these two classifiers, where the weight was set empirically as w=0.6 to weigh the vowel-specific tone classifier more.

Lastly, the final label for each test token is determined from the predicted gender, vowel, and tone class labels.

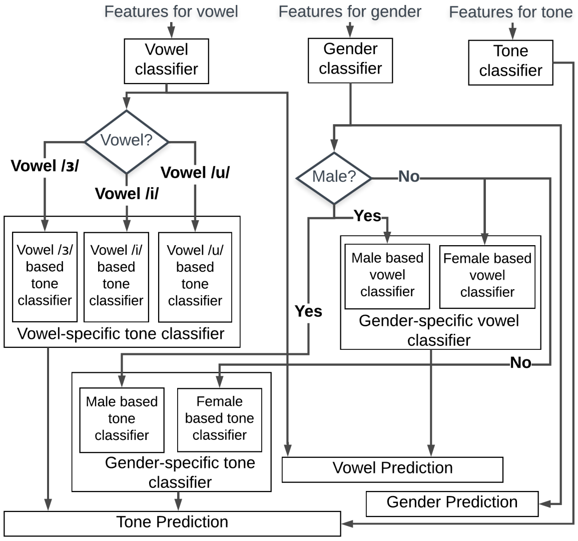
Features for vowel
Features for gender
Features for tone
Vowel classifier
Gender classifier
Tone classifier
Vowel?
Vowel /ɜ/
Vowel /i/
Vowel /u/
Male?
No
Yes
Vowel /ɜ/ based tone classifier
Vowel /i/ based tone classifier
Vowel /u/ based tone classifier
Vowel-specific tone classifier
Male based vowel classifier
Female based vowel classifier
Gender-specific vowel classifier
Yes
No
Male based tone classifier
Female based tone classifier
Gender-specific tone classifier
Vowel Prediction
Gender Prediction
Tone Prediction

Fig. 1: Our proposed hierarchical classification algorithm.

4. EXPERIMENTAL RESULTS & DISCUSSION

In this work, each subcategory for tone, gender and vowel is considered a separate class, thus giving k=24 classes in total. We measure performance using accuracy (ACC) and F1-score (the harmonic mean of precision and recall), but prefer F1-score over ACC [21]. For evaluation, we benchmark the proposed classification framework against those used in previous works[1]:

1. Single multi-class SVM: We trained a single 24-class (and 12-class) SVM that employs *all* features in Table 2.
2. Single multi-class NN: We repeated #1 using the 3-layer neural network (NN) described in [10].
3. Three independent SVMs: We trained 3 SVMs independently for each subcategory, each using its *own* category of features. The independently predicted gender, vowel and tone classes are then used to determine the final labels.
4. Three independent NNs: we repeated #3 with NN ([10]).
5. CNN: We implemented a recent[2] method [7] based on deep-learning and tested it with various settings of hyper-parameters. Results from the best setting are used.

We performed separate sets of cross-validation experiments (CVE) to examine the following questions. Firstly, to examine classification performance under situations where the size of the training set is roughly equal to or larger than that of the test set, we ran t folds such that token samples from only m speakers were used for classification-training while

[1] We chose RBF kernel for SVM to facilitate direct comparisons with related works for Mandarin speech [14] and those for English [22].

[2] Su et al. [7] was designed for pitch determination for a dataset of similar size to ours. We chose this work for comparison over the CNN approach of Qian et al. [23] because the latter required a dataset of over 7000 utterances.

the remaining n-m speakers were used for testing. We set m=$\{10, 15\}$ so that the training data is roughly 50% and 75% of the entire dataset, respectively. The top halves of Tables 3-4 report the results. Shown in each cell are ACC and F1-score obtained by each method for different subcategories and for the overall annotation task. In these tables, we highlight the best performance in blue (cells in pink will be discussed shortly). From these tables, one can see that our proposed algorithm generally achieved the best performance based on both ACC and F1-score, regardless of the subcategory. In contrast, the approach based on CNN [7] performed worse, mostly due to insufficient number of training samples in relation to the number of model parameters employed (e.g. for k=24, ≈92 training samples per class for m=15, and ≈61 per class for m=10). Strategies like data augmentation may mitigate the well known high-dimensionality-low-sample-size problem. Conversely, simpler models like SVM and NN yielded better performance than CNN [7], most likely because fewer model parameters were employed. Lastly, independent classifiers generally performed better than single classifiers, suggesting that the use of category-specific features for each subcategory allowed the individual classifiers to learn class-boundaries within each subcategory better than those for word-level classifications.

Secondly, in pushing the limit of training classification algorithms with small training data, we employed token samples from only 2 randomly selected speakers for training such that each fold has 1 male and 1 female. Thus, each class had ≈23 training samples when k=12, and ≈11 training samples when k=24. We then employed all the samples from the rest of the speakers for testing. We repeated this procedure until all possible combinations of one-speaker-per-gender pairings were examined. From Table 5, one can see that our method maintains the top performance. The achieved F1-score is also significantly higher than that achieved by the competing approach of using independent SVMs, with F1-score for the overall 24-class problem of 0.742 (vs. 0.705).

The last set of CVEs examined how increasing the problem dimensionality to include the annotation of gender may (not) decrease the word-level classification accuracy. We done so by repeating the above CVEs but omitting gender-classification (k=12). Results of these CVEs are shown in the lower halves of Tables 3-5 where we highlighted the cells with highest F1-score for this 12-class problem in pink. Interestingly, the inclusion of gender subcategory has no negative impact on classification performance in all cases. Furthermore, when training data is limited, as in the case of m=2, our approach achieved a significant improvement of 2.8% (F1-score of +0.04) for vowel- classification over the second competing method (3 SVMs), as highlighted in pink in Table 5. In contrast, accuracy of CNN [7] was sub-optimal, with a significant drop in F1-score from 0.810 to 0.661 for vowel-annotation and from 0.679 to 0.451 for tone-annotation for the case of m=10, and likewise for the case of m=2.

Table 3: 4-Fold CV: training data from m=15 speakers. Shown in each cell are accuracies (ACC/F1-score) obtained by each approach for different subcategories and for the overall annotation task. For each subproblem, the cell with the highest F1-score is highlighted in blue (k=24) or pink (k=12).

Approach	Vowel	Tone	Gender	Overall
24-class				
CNN [7]	89.2/0.844	88.1/0.763	85.2/0.813	96.2/0.593
3 SVMs	98.1/0.973	95.9/0.919	97.7/0.972	99.0/0.880
Single SVM	95.5/0.935	95.0/0.900	96.9/0.965	98.6/0.835
3 NNs	97.1/0.957	95.4/0.907	97.7/0.972	98.8/0.857
Single NN	95.7/0.937	94.6/0.892	97.2/0.968	98.6/0.837
Proposed method	98.7/0.981	96.2/0.923	97.7/0.972	99.1/0.892
12-class				
CNN [7]	91.1/0.861	90.3/0.814	NA	NA
3 SVMs	98.1/0.973	95.9/0.919	NA	NA
Single SVM	96.7/0.951	95.4/0.909	NA	NA
3 NNs	96.4/0.946	95.1/0.902	NA	NA
Single NN	95.9/0.939	95.0/0.899	NA	NA

Table 4: 2-Fold CV: training data from m=10 speakers.

Approach	Vowel	Tone	Gender	Overall
24-class				
CNN [7]	85.7/0.795	78.5/0.576	63.0/0.662	94.1/0.368
3 SVMs	96.0/0.941	94.8/0.898	97.5/0.967	98.6/0.850
Single SVM	91.5/0.875	92.8/0.860	94.3/0.923	97.7/0.735
3 NNs	94.2/0.914	94.3/0.886	97.5/0.967	98.3/0.819
Single NN	91.5/0.875	90.8/0.827	95.5/0.933	97.5/0.733
Proposed method	96.3/0.946	95.1/0.902	97.5/0.967	98.7/0.858
12-class				
CNN [7]	86.9/0.810	83.5/0.679	NA	NA
3 SVMs	96.0/0.941	94.8/0.898	NA	NA
Single SVM	93.4/0.905	93.7/0.878	NA	NA
3 NNs	94.5/0.920	94.3/0.887	NA	NA
Single NN	91.9/0.881	92.2/0.849	NA	NA

Table 5: 8-Fold CV: training data from m=2 speakers.

Approach	Vowel	Tone	Gender	Overall
24-class				
CNN [7]	73.2/0.611	71.3/0.431	59.2/0.614	93.2/0.163
3 SVMs	89.9/0.856	92.9/0.858	94.7/0.939	97.5/0.705
Single SVM	83.4/0.758	86.3/0.733	83.3/0.794	95.7/0.510
3 NNs	86.3/0.805	91.4/0.831	94.7/0.939	96.9/0.644
Single NN	81.9/0.732	84.4/0.691	80.6/0.778	95.4/0.459
Proposed method	92.7/0.896	93.1/0.864	94.7/0.939	97.8/0.742
12-class				
CNN [7]	76.4/0.661	72.2/0.451	NA	NA
3 SVMs	89.9/0.856	92.9/0.858	NA	NA
Single SVM	85.0/0.783	89.2/0.786	NA	NA
3 NNs	88.5/0.833	91.7/0.839	NA	NA
Single NN	85.0/0.782	88.1/0.763	NA	NA

5. CONCLUSION

We presented a new hierarchical formulation of AA for Mandarin monosyllabic speech tokens via joint gender-, vowel-, and tone- classification that was shown to better model the inter-dependencies between 3 subcategories than other variants tested. As results show, our approach is advantageous as it did not compromise classification performance even when the complexity of the problem increased while being able to provide additional (gender) information in the annotations. In future, we aim to investigate additional subcategories to our framework, such as more vowels, data sizes and different speech styles, to further aid our target application.

Acknowledgements: This study was funded by research grants from SSHRC and NSERC. We thank Lisa Tang, Keith Leung, Jane Jian, Charles Turo, Dahai Zhang for their help.

References

[1] Lu Zhao et al. “Acoustic features of Mandarin monophthongs by Tibetan speakers”. In: *International Conference on Asian Language Processing*. IEEE. 2014, pp. 147–150.

[2] Puisan Wong. “Acoustic characteristics of three-year-olds’ correct and incorrect monosyllabic Mandarin lexical tone productions”. In: *Journal of Phonetics* 40.1 (2012), pp. 141–151.

[3] S Wang et al. “Lexical tone perception in sensorineural hearing-impaired and auditory neuropathy spectrum disorder”. In: *Journal of Clinical Otorhinolaryngology, Head, and Neck Surgery* 29.17 (2015), pp. 1537–1540.

[4] Keith KW Leung et al. “Acoustic characteristics of clearly spoken English tense and lax vowels”. In: *The Journal of the Acoustical Society of America* 140.1 (2016), pp. 45–58.

[5] Lisa Y.W. Tang et al. “Examining visible articulatory features in clear and plain speech”. In: *Speech Communication* 75 (2015), pp. 1–13.

[6] Andros Tjandra et al. “Combination of two-dimensional cochleogram and spectrogram features for deep learning-based ASR”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Apr. 2015, pp. 4525–4529.

[7] Hong Su et al. “Convolutional neural network for robust pitch determination”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Mar. 2016, pp. 579–583.

[8] Neville Ryant, Jiahong Yuan, and Mark Liberman. “Mandarin tone classification without pitch tracking”. In: *IEEE International Conference on Acoustics Speech and Signal Processing*. 2014, pp. 4868–4872.

[9] Yen-Teh Liu, Yu Tsao, and Ronald Y Chang. “A deep neural network based approach to Mandarin consonant/vowel separation”. In: *IEEE International Conference on Consumer Electronics-Taiwan*. IEEE. 2015, pp. 324–325.

[10] Ozlem Kalinli. “Tone and pitch accent classification using auditory attention cues”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2011, pp. 5208–5211.

[11] Neville Ryant et al. “Highly accurate Mandarin tone classification in the absence of pitch information”. In: *Proceedings of Speech Prosody*. Vol. 7. 2014.

[12] Hongbing Hu et al. “Acoustic features for robust classification of Mandarin tones.” In: *INTERSPEECH*. 2014, pp. 1352–1356.

[13] Qian Liu et al. “A pitch smoothing method for Mandarin tone recognition”. In: *”International Journal of Signal Processing, Image Processing and Pattern Recognition”* 6.4 (2013).

[14] Fuhai Li, Jinwen Ma, and Dezhi Huang. “MFCC and SVM Based Recognition of Chinese Vowels”. In: *Lecture Notes in Artificial Intelligence*. Vol. 3802. 2005, pp. 812–819.

[15] C. C. Chen et al. “Gender-to-Age hierarchical recognition for speech”. In: *2011 IEEE 54th International Midwest Symposium on Circuits and Systems (MWSCAS)*. Aug. 2011, pp. 1–4.

[16] Sarah Ita Levitan, Taniya Mishra, and Srinivas Bangalore. “Automatic Identification of Gender from Speech”. In: *Speech Prosody 2016* (2016), pp. 84–88.

[17] Hsiao-wuen Hon. “A system and method for determining the tone of a syllable of Mandarin Chinese speech”. WO/1996/010248. Apr. 1996.

[18] Dong Enqing et al. “Voice activity detection based on short-time energy and noise spectrum adaptation”. In: *6th International Conference on Signal Processing, 2002*. Vol. 1. Aug. 2002, 464–467 vol.1.

[19] Agnès Belotel-Grenié and Michel Grenié. “The creaky voice phonation and the organisation of Chinese discourse”. In: *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*. 2004.

[20] Hynek Boril and John H. L. Hansen. “Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.6 (Aug. 2010), pp. 1379–1393.

[21] Marina Sokolova and Guy Lapalme. “A systematic analysis of performance measures for classification tasks”. In: *Information Processing & Management* 45.4 (2009), pp. 427–437.

[22] Jeremy Donai, Saeid Motiian, and Gianfranco Doretto. “Automated classification of vowel category and speaker type in the high-frequency spectrum”. In: *Audiology Research* 6.1 (2016).

[23] Yanmin Qian et al. “Very deep convolutional neural networks for noise robust speech recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.12 (2016), pp. 2263–2276.