# IMPROVING FUNDAMENTAL FREQUENCY GENERATION IN EMG-TO-SPEECH CONVERSION USING A QUANTIZATION APPROACH

*Lorenz Diener, Tejas Umesh, Tanja Schultz*

Cognitive Systems Lab, University of Bremen

## ABSTRACT

We present a novel approach to generating fundamental frequency (intonation and voicing) trajectories in an EMG-to-Speech conversion Silent Speech Interface, based on quantizing the EMG-to-$F_0$ mappings target values and thus turning a regression problem into a recognition problem.

We present this method and evaluate its performance with regard to the accuracy of the voicing information obtained as well as the performance in generating plausible intonation trajectories within voiced sections of the signal. To this end, we also present a new measure for overall $F_0$ trajectory plausibility, the trajectory-label accuracy (TLAcc), and compare it with human evaluations.

Our new $F_0$ generation method achieves a significantly better performance than a baseline approach in terms of voicing accuracy, correlation of voiced sections, trajectory-label accuracy and, most importantly, human evaluations.

***Index Terms***— electromyography, EMG-to-Speech, Silent Speech Interfaces, intonation, F0

## 1. INTRODUCTION

*Silent Speech Interfaces* (SSIs) – speech interfaces that continue to function even when an audible acoustic signal is not present – have recently been growing in popularity as a research topic.

SSIs can be built based on many different sensor modalities, such as ultrasound [1, 2], permanent-magnetic articulography [3], microwave radar [4], *surface electromyography* (sEMG, with muscle movement [5] or sub-vocal [6]), non-audible murmur recorded with a throat microphone [7] or even electrocorticography [8].

When built as direct-conversion interfaces – directly converting non-audio biosignals to speech without an intermediate recognition step – SSIs have the potential to replace or augment audible speech. Common suggested use cases of such direct conversion SSIs range from voice prostheses for people who are unable to produce speech (e.g. Laryngectomees) to interfaces for healthy users who may not want to produce audible speech (e.g. in a public place) or want to avoid exhaustion or noise pollution (e.g. agents in a call center).

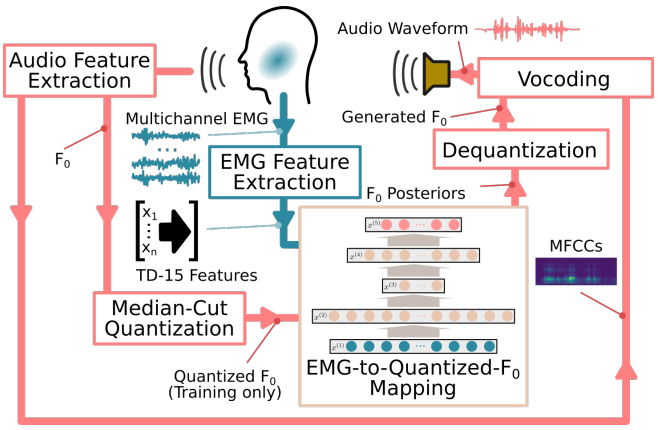


**Fig. 1**. An overview of the system evaluated in this paper, described in Section 2.

While these use cases differ in various ways and present various different challenges, they also have one thing in common: All of them require that the output speech is not only intelligible, but also that it sounds natural – i.e. as much like normal human speech as possible. Most current SSI research, including our own previous work, focuses primarily on intelligibility, due to the large amount of unsolved problems and the difficulty of the challenges involved in building SSIs at all. However, as explained above, this is only one side of the equation – *paralinguistic information* such as intonation carries important context cues that can completely alter the meaning of what is spoken (consider e.g. the sentence *"I never said that she stole my money"*, which can take on seven different meanings depending on where emphasis is placed).

In this work, we present a technique for improving the generation of *fundamental frequency* ($F_0$) trajectories from sEMG data. This technique is based on quantizing the $F_0$ into discrete intervals, turning the task of predicting $F_0$ from a regression into a recognition problem, an approach that has recently been used with some success in audio generation [9].

We present an evaluation of this new technique compared to a baseline regression system introduced by Janke et al. [10] (the state of the art for estimation of $F_0$ from sEMG signals), comparing the voicing accuracy and the quality of generated contours within correctly-recognized-as-voiced sections. We

also present a novel measure for comparing a generated and reference $F_0$ contour and present evidence for the validity of this measure in the form of a subjective listening test evaluation.

The rest of this paper is organized as follows: in Section 2 we describe the EMG-to-Speech conversion systems (the baseline system as well as our new system with $F_0$ quantization). Section 3 presents an overview of the experiments we performed, Section 4 presents their results in terms of common error measures and Section 5 presents our new error measure. The results are discussed in Section 6. Finally, Section 7 will give an outlook on future avenues for research we are currently considering.

## 2. EMG-TO-SPEECH CONVERSION WITH $F_0$ QUANTIZATION

We present two different systems for performing EMG-to-Speech conversion that differ in how they represent $F_0$ contours: once as a sequence of continuous numbers (in the baseline system, labeled "Baseline" throughout the rest of this paper) and once as a set of discrete steps (in our new, proposed quantization based system, labeled "Quantized" in the rest of this paper.). An overview of the structure of our proposed system as evaluated in this paper can be seen in Figure 1.

### 2.1. sEMG input features

The input for the EMG-to-Speech conversion system is, for both of these systems, a set of stacked EMG time-domain features (TD-15 features) [11]. To extract them, we first split each channel of the multi-channel EMG signal into its low (below 134 Hz) and high (above 134 Hz) parts. We then apply windowing, using a 32ms Blackman window with 10ms frame shift (i.e. 22ms overlap between frames).

For each frame, we calculate the power and mean of the low frequency part and the power, rectified mean and zero-crossing rate of the high frequency part. These five features (extracted separately for each channel) together make up the TD feature vector, which is then stacked 15 frames into both the past and the future to create the final TD-15 features which we use as the input for our EMG-to-Speech mapping.

For details on EMG recording and analog signal preprocessing, refer to Section 3.

### 2.2. Audio output features

We extract two different features for the audio signals: the $F_0$, representing excitation and *Mel-Frequency Cepstral Coefficients* [12] (MFCCs), representing the spectral features of the speech signal. For both, the first step is once again to window the signal with a Blackman window and the same frame shift and size as with the sEMG signal.

#### *2.2.1. $F_0$ and quantized $F_0$*

Our system extracts the excitation of the speech signal using the YIN [13] algorithm, a commonly used algorithm for the extraction of speech fundamental frequency trajectories for which good implementations are readily available. This results in one value per frame, giving either the fundamental excitation frequency (in Hz) or 0 for segments of speech that are unvoiced. The trajectories obtained using the YIN algorithm are used to train our models and as a reference for evaluation.

While the baseline system uses regression to directly predict the $F_0$ for each frame, our proposed system converts it to a different representation. For this, it uses the median-cut algorithm [14]. The algorithm works as follows: It starts off considering the entire range of data as one big interval. In each step, the largest remaining interval is then split along its median into two new intervals. This is repeated until the desired number of intervals (32, in our case – determined to be enough to produce good quality output in re-synthesis experiments) is reached. Each interval now represents one class. This process loses very little information: The correlation of voiced sections (compare Section 4.2) between reference $F_0$ value and quantized $F_0$ values is ~0.996 on average.

An $F_0$ value can now be quantized into different classes by looking up into which interval the value would fall. Since the $F_0$ is now split into discrete classes, task of predicting $F_0$ values turns from a regression problem with one continuous output into a recognition problem – predicting one of the 32 $F_0$ classes. While this reduces the resolution available to the $F_0$, it at the same time simplifies the problem, making it easier to train a reliable EMG-to-excitation conversion system. The reverse direction, transforming from classes back to $F_0$ values, can be achieved by looking up the median value for a given class.

As the evaluation data would not be available at training time, we calculate the intervals on the training data of each session and then use these intervals throughout system training and evaluation.

#### *2.2.2. Spectral features and vocoding*

Our system uses 25 MFCCs to model the spectral attributes of the audio signal (the part of the speech signal that is *not* the excitation). These can, together with a $F_0$ trajectory, be converted back into an audible speech waveform using vocoding for evaluation (In this work, we use the AhoCoder [15] vocoder, as it generates more high quality output waveforms compared to the MLSA vocoder used in many previous EMG-to-Speech conversion systems. The parameters and features used have been found to work well for EMG-to-Speech conversion in our previous work [10].

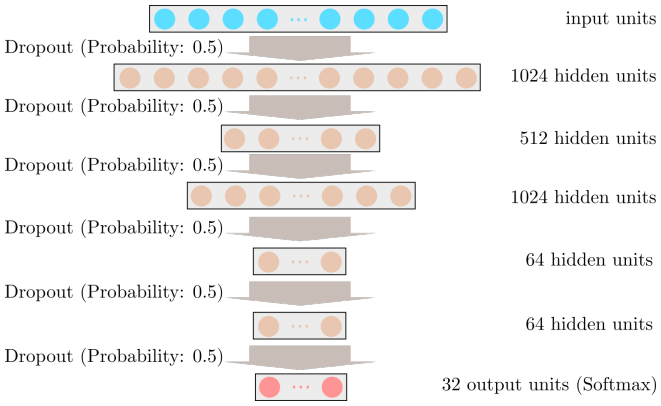


**Fig. 2**. Structure of the feed-forward neural network used for EMG-to-Excitation mapping in this work.

### 2.3. sEMG feature to excitation mapping

To convert EMG signals to $F_0$ trajectories, we use a feed-forward neural network architecture employing a bottleneck shape with ReLU activations, which has been found to work well for this task [16].

#### *2.3.1. Baseline system*

The baseline systems structure is as introduced by Janke et al. [10] (this system, to the best of the authors knowledge, represents the current state of the art for estimation of $F_0$ trajectories from surface EMG signals), with two differences: first, the target features are only the $F_0$ values (we do not consider the mapping of the MFCCs in this work). Second, the parameter optimization is performed using the AdaDelta [17] method with a learning rate of 0.1 and 1024 sample mini-batches, which we have found results in more stable training. When converting sEMG to $F_0$ values, thresholding is applied as a final step, setting all values below 25 Hz to 0 (unvoiced).

#### *2.3.2. Proposed quantized system*

The new proposed system replaces the single linear $F_0$ output unit with a soft-max layer with 32 outputs, one for each possible class in the quantized $F_0$ signal, which is one-hot encoded for training. The system is trained using the AdaDelta method, again with a learning rate of 0.1 and 1024 sample mini-batches. Additionally, since the task to be learned is now a discrete labeling instead of a regression problem, we use a binary cross-entropy loss instead of the mean squared error the baseline uses.

When estimating $F_0$ from sEMG data during evaluation, we always select the class with the highest probability according to the soft-max output for each frame, and then turn it back into a numerical $F_0$ value, as described in Section 2.2.1. An overview of this new structure, which also uses additional layers compared to the baseline system, can be seen in Figure 2. As in the baseline system, ReLU activations are used in all layers except the output layer, which uses a Softmax activation function.

All systems are trained and evaluated in a *session-dependent* manner – they are trained on the training set of only one recording session and then evaluated on the evaluation set of that same session. This is because the sEMG signal, due session differences in exact electrode placement as well as electrode, skin and muscle conditions, is strongly session dependent. Creating session-independent systems, though an area of research that is of significant importance within the SSI community [18, 19], is not the subject of this paper and the data corpus used is not well-suited for such investigations – however, research on session-independent EMG processing is ongoing using publicly available corpus data [20].

## 3. DATA CORPUS AND RECORDING SETUP

To train and evaluate the systems, we use sEMG and audio data recorded in parallel and synchronized with a marker channel present in both signals, and the delay between audio and sEMG signals caused by the physiology of speech production [21] was accounted for by shifting the EMG signal by 50 ms. Note that, since we require audible speech to train and evaluate a system, all evaluations presented are done on sEMG recorded during audible speech (however, the input for the mapping is strictly sEMG data, no audio information is used).

sEMG was recorded using two system setups: One using a setup with six single electrode channels, with the electrodes positioned to capture signals from specific muscles of the articulatory apparatus, and another using two array electrodes (One 4 x 8 matrix electrode positioned on the cheek and one 8 electrode strip positioned below the chin, both with a 10 mm inter-electrode distance and using long-axis-first chained differential derivation). Speech audio was recorded using a standard close-talking microphone in either setup.

**Table 1**. Data Corpus Information

| Speaker-Session | Sex | Length [mm:ss] Train | Dev | Eval | Total. Utts. |
|---|---|---|---|---|---|
| S1-Single | m | 24:23 | 02:47 | 01:19 | 520 |
| S1-Array | m | 28:01 | 03:00 | 00:47 | 510 |
| S1-Array-Lrg | m | 68:56 | 07:41 | 00:48 | 1103 |
| S2-Single | m | 24:12 | 02:42 | 00:49 | 509 |
| S2-Array | m | 22:14 | 02:25 | 01:10 | 520 |
| S3-Array-Lrg | f | 110:46 | 11:53 | 00:46 | 1977 |
| **Total** | | 278:32 | 30:28 | 05:39 | 5139 |

The corpus we use contains a total of six sessions recorded by three speakers (two male, one female) using these setups.

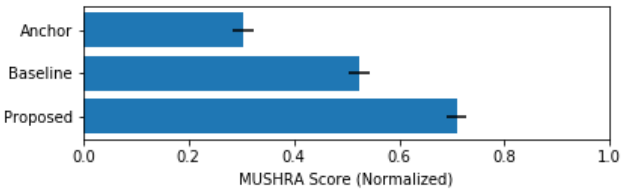


**Fig. 3**. Results of the MUSHRA listening test, showing the scores for the anchor, the baseline system and our proposed quantization-based system. Scores normalized by reference score to account for inter-rater differences. Higher is better, error bars indicate 95% confidence interval.

Each recording consists of a set of phonetically balanced English sentences from the broadcast news domain (For the large sessions, this was additionally augmented with sentences from the CMU Arctic [22] and TIMIT [23] corpora). In order to ensure consistent pronunciation of words, recordings were supervised by a researcher and participants were allowed to re-record sentences if they or the recording supervisor deemed the recording quality unsatisfactory.

The data was split into a training, development and evaluation set. The development set was used during explorative research and hyper-parameter optimization, whereas the evaluation holdout was used to produce final figures for this paper. A detailed breakdown of the data can be found in Table 1.

## 4. EVALUATION

### 4.1. Subjective listening test

We evaluate the performance of the baseline and our proposed systems using the *MUltiple Stimuli with Hidden Reference and Anchor* (MUSHRA) [24] method, implemented by the BeaqleJS [25] framework. We have each participant of the listening test listen to – and rate the naturalness of – four audio files generated from 30 utterances each (5 randomly selected utterances from the evaluation set of each session), synthesized using the reference MFCCs and the $F_0$ from either the baseline system (audio 1) or our proposed system (audio 2). A completely flat $F_0$ trajectory (the median of the reference $F_0$ trajectory for each utterance), resulting in a robotic sound, is used as a low anchor (audio 3), and the re-synthesized signal (using the reference MFCCs and $F_0$ trajectory) is used as the reference (audio 4). Both the order of the utterances within the test and the order of options for each utterance were randomized. A total of 21 people participated in the listening test, one of which we removed from the results because they consistently rated the reference as 0 points, indicating that they did not complete the test as instructed. This leaves us with data from 20 participants for each of the 30 utterances evaluated for a total of 600 ratings.

To account for inter-rater differences in speech naturalness perception, we normalize the ratings according to the rating assigned to the reference in the MUSHRA test. The resulting scores can be seen in Figure 3. It is clear that the proposed system is considered significantly (verified using a two-tailed independent sample t-test not assuming equal variance at a level of $p < 0.05$) better than the baseline system by the listening test participants.

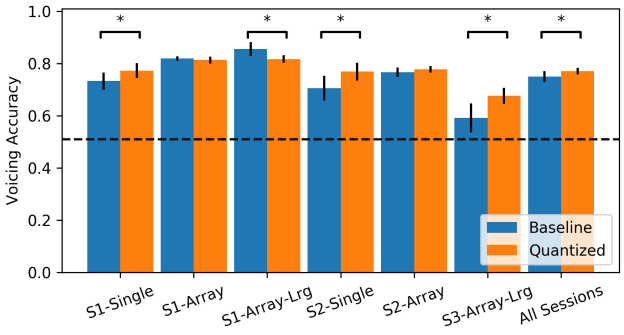


**Fig. 4**. Voicing accuracy of the $F_0$ trajectories generated by the baseline and our proposed quantization-based system. Higher is better, error bars indicate 95% confidence interval, dashed line indicates chance level, '*' indicates significant differences.

### 4.2. Objective evaluation

We further evaluate the performance of our proposed method compared to the baseline with two objective measures capturing different aspects of the $F_0$ signal.

The first measure is the voicing accuracy, i.e. the ratio of frames for which a system has correctly assigned either the value 0 (i.e. the frame is unvoiced) or a value other than 0 (i.e. the frame is voiced), compared to the reference $F_0$ trajectory. The results of this evaluation can be seen in Figure 4. Note that the chance level in the data set would be ~0.51, as voiced and unvoiced or silent frames are present in roughly equal measure.

It can be seen that while our system significantly improves the accuracy for some of the sessions – specifically, sessions S1-Single, S2-Single and S3-Array-Lrg (tested using a paired two-tailed t-test at $p < 0.05$), it actually performs significantly worse for one (S1-Array-Lrg), and for the two remaining sessions there are no statistically significant differences between the two systems outputs. Averaged over all sessions, the baseline system achieves an accuracy of ~0.75, whereas our proposed system achieves an accuracy of ~0.77 – this improvement is, once again, significant.

The second measure that we consider looks specifically at the voiced sections of the signal: We calculate the correlation of the hypothesized and the reference $F_0$ signals, restricted to only those parts that are voiced in the reference and were correctly recognized as voiced (i.e. the $F_0$ value is not 0) in

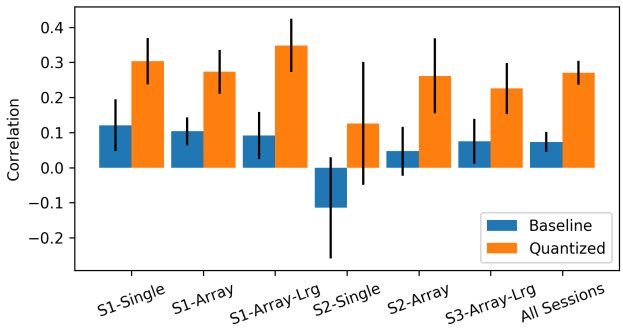


**Fig. 5**. Correlation of the $F_0$ trajectories generated by the baseline and our proposed quantization-based system, calculated only within segments correctly considered as voiced by the systems. Higher is better, error bars indicate 95% confidence interval.

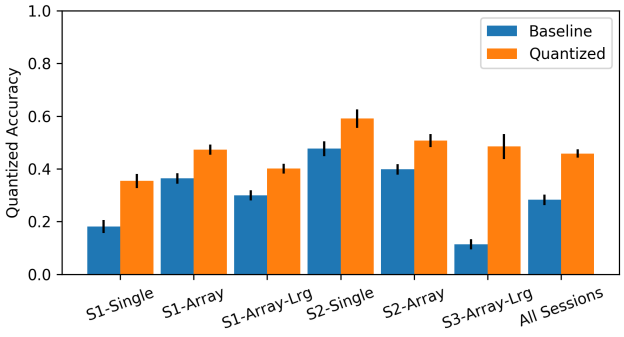


**Fig. 6**. Accuracy of the quantized $F_0$ values of the baseline and our proposed system. Higher is better, error bars indicate 95% confidence interval.

the systems output. The correlations obtained in this way for each session can be found in Figure 5. It is clear that while the baseline system can discriminate voiced and unvoiced sections, its performance with regard to generating a good $F_0$ trajectory seems to be severely lacking (average correlation of ~0.07). Our proposed system, on the other hand, produces significantly better (tested at $p < 0.05$ with a two-tailed paired t-test, overall as well as within each session) results (average correlation: ~0.27).

It should be noted that, while the improvement is significant, the correlations are very low especially for the baseline system. This should not come as a surprise, as the task of estimating $F_0$ from surface EMG data is hard – the base excitation of the voice box is nor measured by facial EMG, so only indirect inference is possible. It is for this reason that we introduce a measure which we believe to be better suited for the comparison of $F_0$ trajectories especially in settings where the output quality of the systems being compared is relatively low in the following section.

Finally, we consider the performance of the recognizer for the quantized $F_0$ and compare it with the baseline system by taking that systems output, quantizing in the same way as we did for the quantized system (compare Section 2.3.2) and then calculating the recognition rate against the reference $F_0$ trajectory. The results can be seen in Figure 6 – the proposed system significantly (tested at $p < 0.05$ with a two-tailed paired t-test) outperforms the baseline.

## 5. TRAJECTORY-LABEL ACCURACY

While the measures mentioned above quantify the quality of the systems output objectively, the correlation is somewhat hard to interpret and neither measure shows the whole picture – our system may have performed worse with regard to voicing accuracy for the sessions where this accuracy was high to begin with, but is this offset by performing better in voiced sections?

To ameliorate the issues we see with these measures, we introduce a new objective measure for comparing generated $F_0$ trajectories to a reference: The *trajectory-label accuracy* (TLAcc). Intuitively, it combines both the accuracy of the voicing as well as how well voiced sections are restored, reducing the movement within voiced sections to its most basic components: going up, going down, or neither of the two. While this is a major simplification of the complexity of $F_0$ movement during speech, this allows comparisons of whether a generated trajectory is basically similar to another – unlike other representations or annotation schemes, which may provide too much detail to allow for meaningful comparisons, or may not even be accurately extractable without human assistance.

We are able to show that, on our results, this measure is more strongly correlated with subjective assessments of naturalness than voicing accuracy or correlation while still being easy to evaluate without human interference, potentially making it a better candidate for use during system development than those measures.

### 5.1. Calculating the TLAcc

The TLAcc is calculated as follows: First, the central differences gradient of the $F_0$ trajectory is calculated by subtracting the value of the frame right of the current frame from the value left of the current frame – however, if either the value of the frame to the left or to the right of the current frame is zero (unvoiced), the central value is used instead of that value. Then, labels are assigned:

- ”unvoiced” (the $F_0$ value of this frame is zero / unvoiced),
- ”rising” (the $F_0$ value rises by at least 5 Hz, according to the calculated gradient)
- ”falling” (the $F_0$ value falls by at least 5 Hz, according to the calculated gradient)
- ”flat” (otherwise)

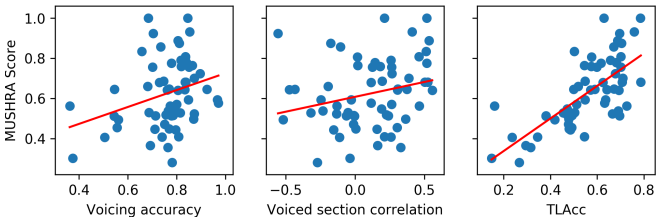


**Fig. 7**. Scatter plots of utterance mean normalized MUSHRA scores (baseline and quantized systems) against the ratings assigned to the same utterances by the voicing accuracy (left), the voiced section correlation (middle) and our proposed trajectory-label accuracy measure (right). Red lines indicate regression line.

The trajectory-label accuracy is then the accuracy calculated between the reference and hypothesis trajectory labels. A reference python implementation of this measure is available online[1].

To verify that this is a sensible approach, we compare the ratings produced by this new measure to the human evaluations from the subjective listening test (compare Section 4.1).

A scatter plot of the MUSHRA scores (average per utterance scores) versus the three different objective measures for all utterances from the listening test can be seen in Figure 7, showing that the trajectory label accuracy is strongly correlated with the listening test scores (Pearson's $r \approx 0.71$, compared to only $r \approx 0.3$ for the voicing accuracy and $r \approx 0.25$ for the voiced section correlation).

### 5.2. Evaluation of proposed method using TLAcc

Finally, we perform an evaluation of the baseline and our proposed system according to trajectory-label accuracy, which can be found in Figure 8. In this, our proposed quantization-based system clearly and significantly (tested at $p < 0.05$ with a two-tailed paired t-test) outperforms the baseline approach (average trajectory label accuracies: ~0.42 for the baseline system compared to ~0.59 for the proposed approach).

## 6. DISCUSSION

Paralinguistic attributes of speech such as intonation should be an important consideration in trying to build direct synthesis Silent Speech Interfaces, yet they are often only considered on the side. One reason for this may be that the quality of synthesized intonation is hard to measure – a $F_0$ contour may differ from the reference to some extent, but still be perfectly sensible.

In this paper, we have made two contributions: (1) we have presented a novel recognition-based prediction of $F_0$ which outperformed the baseline system for direct synthesis SSIs, and (2) we have introduced the TLAcc measure which

[1] https://github.com/cognitive-systems-lab/trajectory-label-accuracy

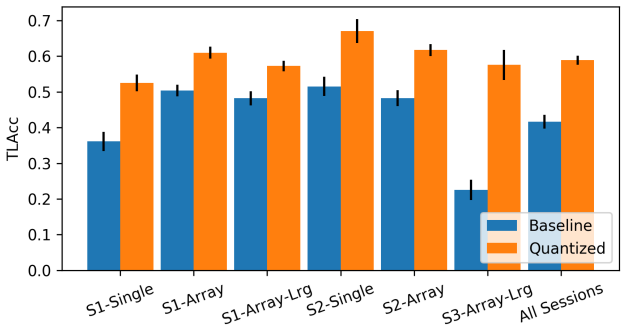


**Fig. 8**. Our proposed trajectory-label accuracy measure for the baseline and our proposed quantization-based system. Higher is better, error bars indicate 95% confidence interval.

correlates well with human listening tests and thus allows to better quantify improvements than traditional measures.

In Section 4.1, we presented an evaluation of our new proposed method according to the gold standard in evaluating speech audio – human listening tests. This clearly demonstrates that our proposed method manages to improve upon the baseline system, however, such listening tests are time-consuming and are only sensible as a final evaluation and not during development and tuning of new methods.

For this reason, we considered some objective measures that can be quickly and automatically calculated in Section 4.2, again demonstrating that our new system improves upon the baseline to some extent (and, when considering the quantized accuracy, the measure that the proposed system optimizes directly, to a large extent – however, such an evaluation seems unfairly biased towards the proposed system). To address the shortcomings we see with these measures, in Section 5, we introduced a new objective measure – the TLAcc – and were able to show that it correlates well (and better than other objective measures) with human listening test scores.

## 7. SUMMARY

We have presented a novel approach that can be used to improve excitation predictions of direct-synthesis SSIs. This approach significantly outperforms a baseline regression approach in human listening test evaluations as well as, to some extent, on objective measures.

We have additionally introduced the TLAcc, a new objective measure of $F_0$ trajectory quality which we hope will be useful in easing the development future SSI systems with regard to the quality of the generated intonation.

In the future, we may further explore quantization based approaches to direct-synthesis SSI, potentially taking a quantize-and-recognizes approach for not only the excitation, but also for the spectral features and further experimenting with the amount of intervals used for quantization.

## 8. REFERENCES


[1] Tamás Grósz, Gábor Gosztolya, László Tóth, Tamás Gábor Csapó, and Alexandra Markó, "F0 estimation for dnn-based ultrasound silent speech interfaces," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 291–295.

[2] Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda, and Pierre Badin, "Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract," *Speech Communication*, vol. 93, pp. 63–75, 2017.

[3] Jose A Gonzalez, Lam A Cheah, James M Gilbert, Jie Bai, Stephen R Ell, Phil D Green, and Roger K Moore, "A silent speech system based on permanent magnet articulography and direct synthesis," *Computer Speech & Language*, vol. 39, pp. 67–87, 2016.

[4] Peter Birkholz, Simon Stone, Klaus Wolf, Dirk Plettemeier, Klaus Wolf, and Dirk Plettemeier, "Non-invasive silent phoneme recognition using microwave signals," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 12, pp. 2404–2411, 2018.

[5] Lorenz Diener and Tanja Schultz, "Investigating Objective Intelligibility in Real-Time EMG-to-Speech Conversion," in *INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association*, 2018.

[6] Arnav Kapur, Shreyas Kapur, and Pattie Maes, "Alterego: A personalized wearable silent speech interface," in *23rd International Conference on Intelligent User Interfaces*. ACM, 2018, pp. 43–53.

[7] Tomoki Toda and Kiyohiro Shikano, "Nam-to-speech conversion with gaussian mixture models," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2005, pp. 1957–1960.

[8] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ecog: A pilot study," in *Engineering in Medicine and Biology Society (EMBC), 2016 38th Annual International Conference of the IEEE*, 2016.

[9] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[10] Matthias Janke and Lorenz Diener, "Emg-to-speech: Direct generation of speech from facial electromyographic signals," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 25, no. 12, pp. 2375–2385, nov 2017.

[11] Szu-Chen (Stan) Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel, "Towards continuous speech recognition using surface electromyography," in *In proceedings of the 9th ISCA International Conference on Spoken Language Processing*, 2006.

[12] Satoshi Imai, "Cepstral analysis synthesis on the mel frequency scale," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1983, vol. 8, pp. 93–96.

[13] Alain De Cheveigné and Hideki Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[14] Paul Heckbert, *Color image quantization for frame buffer display*, vol. 16, ACM, 1982.

[15] Daniel Erro, Inaki Sainz, Eva Navas, and Inma Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[16] Lorenz Diener, Matthias Janke, and Tanja Schultz, "Direct conversion from facial myoelectric signals to speech using deep neural networks," in *International Joint Conference on Neural Networks*, 2015, pp. 1–7.

[17] Matthew D Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701*, 2012.

[18] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel, "Session independent non-audible speech recognition using surface electromyography," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, 2005.

[19] Michael Wand and Tanja Schultz, "Session-independent emg-based speech recognition," in *International Conference on Bio-inspired Systems and Signal Processing*, 2011.

[20] Michael Wand, Matthias Janke, and Tanja Schultz, "The emg-uka corpus for electromyographic speech processing," in *The 15th Annual Conference of the International Speech Communication Association, Singapore*, 2014, Interspeech 2014.

[21] PR Cavanagh and PV Komi, "Electromechanical delay in human skeletal muscle under concentric and eccentric

contractions," *European Journal of Applied Physiology and Occupational Physiology*, vol. 42, no. 3, pp. 159–163, 1979.

[22] John Kominek and Alan W Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[23] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[24] ITU Radiocommunication Bureau, "Method for the subjective assessment of intermediate quality level of coding systems," *Recommendation ITU-R BS. 1534*, 2001.

[25] Sebastian Kraft and Udo Zölzer, "Beaqlejs: Html5 and javascript based framework for the subjective evaluation of audio quality," in *Linux Audio Conference, Karlsruhe, DE*, 2014.