Voice expression conversion with factorised HMM-TTS models

Javier Latorre, Vincent Wan, Kayoko Yanagisawa
Toshiba Research Europe Ltd., Cambridge Research Lab, 208 Science Park, Cambridge, UK.
{javier.latorre, vincent.wan, kayoko.yanagisawa}@crl.toshiba.co.uk

Abstract
This paper proposes a method to modify the expression or emotion in a sample of speech without altering the speaker’s identity. The method exploits a statistical speech model that factorises the speaker identity from expressions using linear transformations. For this approach, the set of transforms that best fit the speaker and expression of the input speech sample are learned. They are then combined with the expression transformations of the desired expression taken from another speaker. Since the combined expression transform is factorised and contains information about expression only, it may be applied to the original speech sample to modify its expression to the desired without altering the identity of the speaker. Notably, this method may be applied universally to any voice without the need for a parallel training corpus.

Index Terms: Speech synthesis, voice conversion, expression conversion, parametric speech synthesis

1. Introduction
Two of the main challenges of speech synthesis are in improving audio quality and controllability. Traditionally, these two goals have been considered to be contradictory. In most cases, high quality implies that the system can produce only a very limited variety of voices or styles, whereas high controllability is achieved at the expense of quality. The two main streams for speech synthesis technology, unit selection and parametric speech synthesis, reflect this division.

Unit selection systems produce speech by sampling and concatenating real audio units from a large database so that the audio needs very little modification, if at all. In this way, units can preserve their original quality of the recorded speech. The main problem is in the concatenation when adjacent units do not merge well. To minimise this problem, the concatenated units should be as homogeneous as possible in terms of expression, style, etc. Given the high cost that recording a large database for each type of voice would imply, in practice unit selection synthesis is only aimed at achieving very high quality in a single neutral voice.

On the other hand, statistical parametric speech synthesis does not deal with the audio samples directly, but with their statistical representation. By modifying the distributions it is possible to obtain an enormous degree of controllability in terms of speaker voice, expression, language, etc [1, 2]. The price to be paid for being able to build such statistical models is that the audio signal must be parameterized. This results in a degradation of the audio quality w.r.t. recorded speech. There is also a third option that uses voice conversion (VC). In this case, the ideal is to have a unit selection system as the basis, and apply some VC techniques to generate synthetic speech with the desired speaker voice, expression, etc.

For most approaches, VC consists of learning the mapping between the acoustic features of a source and a target speaker. Multiple methods have been proposed to learn that mapping, such as look-up table [3], GMM [4, 5], neural networks [6] or Gaussian process [7]. Their main disadvantage is that to learn that mapping a parallel corpus between the source and target speaker is needed, and acquiring it is again problematic. Other approaches [8–10] alleviate this requirement by using the GMM trained from multiple parallel corpora of other reference speakers, and projecting the source and target speakers into the source and target speaker-spaces obtained from those other corpora. Finally, model-based VC completely removes the need for parallel corpora. Instead, it uses the concept of bringing the source and target speaker voices into a common space using a statistical model and a set of linear transforms [11–13]. Converting voice A into voice B can be achieved by combining the transforms that map A into the common space with the one that recovers B from that common space.

Many of these techniques have also been used to convert neutral speech into expressive speech, for spectrum, prosody or both [14–20]. All these expression conversion proposals are trained on a parallel corpus containing neutral and expressive speech, usually uttered by a single speaker. Therefore, the transformation can only be safely applied to modify the expression of neutral speech for that specific speaker.

This paper proposes a novel technique to transform only one aspect, the expression, of the input voice while leaving the speaker’s identity unmodified. It is a model-based approach that also provides an integrated framework to modify the spectrum, prosody and other high-level suprasegmental properties without using any parallel corpus. Two branches of the technique are explored: one based on an expression transform learned from a reference speaker; the other based on speaker and expression factorisation.

The rest of the paper is as follows. Section 2 describes the proposed method. Section 3 describes the experiments and results and finally conclusions are drawn in Section 4.

2. Model-based expression conversion
Let the input data from a target speaker s with expression n be called \(\mathbf{o}_{sn}\). The goal of the transformation is to find a function \(f(\mathbf{o}_{sn}, e) = \mathbf{o}_{se}\) that transforms expression n into expression e without modifying the speaker identity. In a statistical framework this function is expressed as

\[
f(\mathbf{o}_{sn}, e) = \arg\max_{\mathbf{o}_{se}} P(\mathbf{o}_{se} | \mathbf{o}_{sn}, e, \Lambda)
\]

where \(\Lambda\) is the voice conversion model. In mapping-based approaches, \(\Lambda\) is trained on parallel data for both the input and the target expressions. Model-based voice conversion does not require parallel data. Instead, it assumes the existence of a canonical space \(\hat{o}\) independent of speaker and expression that can be
obtained from the observed space by means of linear transforms \( \{A, b\} \) as
\[
\hat{o} \approx A_{sn} o_{sn} + b_{sn} \approx A_{se} o_{se} + b_{se} \approx A_{sy} o_{sy} + b_{sy} \tag{2}
\]
These transforms can be factorised as a combination of speaker and expression transforms so that
\[
A_{sn} o_{sn} + b_{sn} = A_s (A_n o_{sn} + b_n) + b_s \tag{3}
\]
Identifying terms
\[
A_{sn} = A_s A_n \tag{4}
\]
\[
b_{sn} = A_s b_n + b_s \tag{5}
\]
If \( A_s \) and \( A_n \) are invertible, the original non-canonical space can be recovered as
\[
o_{se} = A_s^{-1} (\hat{o} - b_{se}) + \epsilon \tag{6}
\]
with \( \epsilon \) a noise term. Substituting equations (2) to (5) into (6) gives
\[
o_{se} = A_s^{-1} A_{sn} -1 (\hat{o} - A_s A_n + b_s) + b_s - \hat{o} + b_{se} + \epsilon = A_s^{-1} A_n o_{sn} + A_{se} (b_n - b_s) + \epsilon. \tag{7}
\]
Assuming \( \epsilon \) to follow a Gaussian distribution with zero mean and \( \sigma_e \) variance and
\[
o_{sn} \sim \mathcal{N}(o_{se}; \mu_{sn}, \Sigma_{sn}) \tag{8}
\]
then
\[
P(o_{se}|o_{sn}, \epsilon, A) = \mathcal{N}(o_{se}; \mu_{se}, \Sigma_{se}), \tag{9}
\]
where
\[
\mu_{se} = A_s^{-1} A_n o_{sn} + A_{se} (b_n - b_s) \tag{10}
\]
\[
\Sigma_{se} = (A_s^{-1} A_n) \Sigma_{sn} (A_s^{-1} A_n)^\top + \sigma_e I, \tag{11}
\]
with \( I \) the identity matrix. For simplicity, the term \( \sigma I \) is assumed to be small and may be dropped. If the space \( o \) consists only of static coefficients \( c \), the output would be directly \( c_{se} = \mu_{se} \). However, this is likely to introduce discontinuities. To avoid them, \( o \) can include dynamic features so that \( o = We \). Then, a smooth transform set of parameters can be obtained using maximum-likelihood parameter generation (MLPG) [21] as
\[
c_{se} = (W^\top \Sigma_{se}^{-1} W)^{-1} W^\top \Sigma_{se}^{-1} \mu_{se}. \tag{12}
\]
If no duration modification is required, then \( \mu_{sn} = o_{sn} \). It should be observed that in that case if no transform is applied, (12) yields \( c_{se} = o_{sn} \).

Obtaining the covariance \( \Sigma_{sn} \) out of \( o_{sn} \) can be problematic. Fortunately, MLPG is robust against variance modifications [22]. Therefore, \( \Sigma_{sn} \) can be obtained from the canonical model covariance \( \Sigma \) as \( \Sigma_{sn} = A_{sn} \Sigma A_{sn}^\top \). Alternatively, \( \Sigma_{se} \) could be obtained directly as
\[
\Sigma_{se} = A_{se}^{-1} A_{sn} A_{sn}^{-1} A_{se}^\top. \tag{13}
\]

2.1. Duration modification

The above description assumed a single linear transformation between the observed and the canonical space. Although this is not true over the whole space, a piece-wise linear approximation over limited regions is still valid. For that purpose, transforms are context dependent. If the sequence of context is unknown, then each frame \( t \) in (9) may be estimated by marginalising over all possible contexts \( m \) as
\[
P(o_{se}(t)|o_{sn}(t), A) = \sum_{m=1}^M P(m|o_{sn}(t), A) \mathcal{N}(o_{se}; \mu_{se}, \Sigma_{se}), \tag{14}
\]
which is basically the same as the equation for the GMM-based conversion method [4].

However, it can be assumed that the text and the phonetic sequence associated with the input speech is known [12, 23] and the transform to be applied depends only on the context, the specific transform \( m \) to be applied at each frame \( t \) can be determined using a forced Viterbi alignment. This approach also allows for easy modification of the speech duration, which is an important component for some expressions [24–26]. A modified duration can be achieved by resampling each segment \( h \) of input speech with a linear transform
\[
m_{sn}^{e,h} = M_{s,h(n)} o_{sn}^{h,n}, \tag{15}
\]
where \( M_{s,h(n)} \) is a resampling matrix with dimensions \((d_{sn}^h, d_{sn}^h)\) in which \( d_{sn}^h \) is the expected duration of segment \( h \) for speaker \( s \) in the intended expression \( e \), and \( d_{sn}^h \) is the original dimension of \( o_{sn}^h \). In the simplest case \( M_{s,h(n)}^{h,n} \) can be defined to yield the mean value of each input segment so that \( \forall t \in h \)
\[
m_{sn}^{e,h}(t) = \frac{\sum_{\tau \in h} o_{sn}^h(\tau)}{d_{sn}^h}. \tag{16}
\]

The target duration distributions can be obtained from a speaker- and expression-independent duration model using a set of factorised linear transforms \( \{\alpha, \beta\} \) similar to (9)
\[
P(d_{sn}^h|d_{sn}^h, A) = \mathcal{N}(d_{sn}^h; \alpha_e d_{sn}^h + \beta_e d_{sn}^h - \beta_e, \alpha_e^{-2}). \tag{17}
\]
Again, \( \sigma_d \) can be hard to obtain from \( o_{sn} \), so it is better to compute it by transforming the variance of the canonical model \( \sigma \) as \( \sigma_{sn} = \sigma \alpha_{sn}^{-2} \) or alternatively directly compute \( \sigma_{se} = \sigma \alpha_{se}^{-2} \).

2.2. Expression conversion with non-factorised models

In the case of factorised models [1, 27], the expression transforms \( \{A_s, b_s\} \) and speaker transform \( \{A_n, b_n\} \) can be obtained directly from the input and the training data. However, expression transformation is also possible for non-factorised models as long as the training data includes one speaker \( n \) in an expression \( n \) sufficiently close to the input expression \( n \) to safely assume \( A_n \approx A_{sn} \) and \( b_n \approx b_{sn} \). In that case, the terms in (7) become
\[
A_{se}^{-1} A_n \approx A_{se}^{-1} A_{sn} = A_{se}^{-1} A_{sn} = A_{se}^{-1} A_{sn}. \tag{18}
\]
and, with \( b_{sn} - b_e = A_r(b_{sn} - b_e) \) from (5)
\[
A_{se}^{-1} (b_n - b_e) = A_{se}^{-1} A_{se} A_r (b_n - b_e) \tag{19}
\]
\[
\approx A_{re} (b_{sn} - b_{se}). \tag{20}
\]
Consequently, the target space can be approximated as
\[
o_{se} \approx A_{re}^{-1} (A_{se} o_{sn} + (b_{se} - b_{se})). \tag{21}
\]
2.3. Expression conversion with factorised CAT models

Cluster adaptive trained (CAT) models can be considered a special form of CMLLR in which all matrices \( A \) are identity and the \( b \) vectors are given by the summation of different mean vectors each provided by an independent decision tree (as described in [28]). An advantage of the CAT structure is that speaker and expression factors can be distributed into different clusters thereby providing a factorised model. Additionally the expression clusters may be interpreted as a set of bases that define a continuous expression space that can produce more complex expressions that are interpolations or extrapolations of the in-training expressions. This gives rise to voice expression conversion with a high level of controllability.

This allows the speaker and expression transforms of a context \( m \) to be defined as

\[
b_m = -\sum_{i=1}^{P_s} \mu_{m,i} \lambda_{i,m}^{(s)}
\]

\[
b_m = -\sum_{i=1}^{P_e} \mu_{m,i} \lambda_{i,m}^{(e)}
\]

where \( P_s \) and \( P_e \) are the number of clusters associated with speaker and expression factors respectively, \( \mu_{m,i}^{(s)} \) and \( \lambda_{i,m}^{(e)} \) are the mean vector and CAT weights for context \( m \) for the \( i^{th} \) speaker or expression cluster.

Implementing voice modification within this framework requires a substitution of the bias cluster of the original CAT model with \( b_{sn} \), setting the speaker weight vector \( \lambda_{m}^{(sn)} \) to zero, and defining the expression weight vector as the difference between the target and input expression vectors \( \lambda_{m}^{(n)} - \lambda_{m}^{(s)} \).

3. Experiments

To the best of our knowledge, no other technique presently allows voice transformation of both spectrum and prosody components with the aim of modifying the expression while retaining speaker identity without recourse to a parallel corpus. In related work on speaker and emotion factorisation using CAT [1], prosody transplant was used as baseline but there CAT produced the desired expressions better. Running an ABX comparison against a known weak baseline system that only modifies the prosody is uninformative.

 Whilst acknowledging the lack of a proper baseline, the goal of the evaluation is to determine to what degree the proposed technique achieves its goal. For this purpose two sets of DMOS tests were designed: one to test whether the system still keeps the speaker similarity after applying the expression conversion; and the other to test the similarity between the modified speech and the intended expression.

The speech is vocoded using 40 Mel-cepstral coefficients including energy, log-F0 and 21 bark-scaled aperiodicity bands with their first and second order deltas.

3.1. Expression model

The CAT model on which both factorised and non-factorised conversion is based was trained in a similar way as the one described in [1]. It has 4 clusters to represent speakers and 3 clusters to represent expressions. The main difference is that here the expression CAT weights of different speakers are not tied for expressions other than neutral in order to avoid the problems due to different renditions of the same expression (as described in [1]). The model was trained on 4 speakers and 9 expressions including neutral, angry, happy, sad, surprised, fast, slow, formal and informal (although only two of the speakers had expressive data).

3.2. Input data

The data of two input speakers was used in these experiments: one male and one female. For each speaker, 90 sentences were used for adapting the CAT model’s weight vectors to each input speaker (although CAT supports adaptation with far fewer sentences). A further 9 sentences per input speaker were used to test the expression conversion. Each of these sentences were aligned at the state-level with the CAT model and the corresponding adapted weight vectors. These alignments were then used to create pseudo-models for the CAT bias cluster. Pseudo-models are individual context-dependent models that have their means substituted by the median observation vector on a state-by-state basis. All the other clusters were copied directly from the CAT model yielding a complete CAT pseudo-model. The median was used instead of the mean because the number of frames in each state is usually small. Therefore, the median provides a more robust estimation of the real mean. For similar reasons, instead of computing the variance from the aligned samples, the variance of the CAT model was used. For the conversion, the expression weights from the expressive male in-training speaker were applied to the male input voice and those of the expressive female in-training speaker to the female input voice. These test sentences were modified into three different expressions: angry, happy and sad.

3.3. Global variance factorisation

Since the proposed conversion is based on parameter generation it might suffer the same over-smoothing problems as standard HMM-based synthesis. Global variance (GV) [29] can alleviate this problem. However, GV is also expression and speaker dependent. Fortunately, GV does not require much data to be computed. Therefore, the adaptation data used to obtain the expression CAT weights for the testing speakers were also used to create their GV models. Then, these models were transformed to the target expression with the non-factorised approach. Here, instead of CAT-based GV, a simpler model interpolation was used: the mean and variance of the transformed GV models were defined as

\[
\mu_{se}^{GV} = \mu_{re}^{GV} + \kappa (\mu_{rn}^{GV} - \mu_{re}^{GV})
\]

\[
\Sigma_{se}^{GV} = \Sigma_{rn}^{GV} + \kappa^2 (\Sigma_{rn}^{GV} - \Sigma_{re}^{GV})
\]

with \( \kappa \) the level of expression conversion which, for this experiment, was always 1.0.

3.4. Subjective tests

Subjective DMOS listening tests were crowdsourced [30] via CrowdFlower to evaluate the similarity of the speaker and of the expression. All stimuli were judged by 10 subjects. To test speaker similarity, subjects were played the original unmodified, unvocoded sample of the speech followed by the modified sample. They were then asked to rate the similarity of the speakers in the two samples ignoring differences in quality or expression as much as possible. In expression similarity tests, subjects heard the modified sample along with a non-vocoded sample of a different sentence but in the same expression; that reference sentence was taken from the corresponding in-training speaker of the same gender. In total 68 subjects participated on the speaker similarity test and 94 in the expression similarity test.
3.5. Speaker similarity

The upper anchor (PM-NEU) for this experiment is the speech generated by the CAT pseudo-models without any form of modification, i.e. $b_s = 0$. The lower anchor is the expressive speech generated by the CAT-TTS system (TTS); the CAT weights are those obtained by adapting to the input data with the expression components of those weights modified to be the difference between the target and adapted-input expression weights (see section 2.3). The proposed systems are labelled VCF and VCNF for the factorised and non-factorised version respectively. As an additional reference, the TTS samples from the speaker adapted CAT model without any expression modification were also included (TTS-NEU).

Figure 1 shows the results in a boxplot. The dots and numbers in each box indicate the mean score. As expected, the proposed methods achieve significantly better speaker similarity than TTS. However, the similarity is significantly worse than that of PM-NEU due to the expression modification. It seems that the speaker similarity degradation introduced by the expression conversion is equivalent to that introduced by the CAT adaptation.

3.6. Expression similarity test

The upper anchor is given by the pseudo-models built on a further 9 utterances from each of the two in-training speakers for each expression. The expression CAT weights for these reference-speaker pseudo-models must be set to 0 as the expression is already part of the bias. The reference-speaker pseudo-models are labelled PM-REF. The lower anchor is the pseudo-model built on the target speaker without expression modification (PM-NEU). As above, this test also includes the samples from the speaker adapted CAT model in the intended expression (TTS).

Figure 2 shows the boxplot of the results for each emotion. The expression similarity of the proposed methods are between that of the unmodified input speech and the intended expression. Therefore it can be concluded that the proposed method is effective in changing the speech expression, though there is still substantial room for improvement. In general, the expression similarity of the proposed voice conversion systems is equivalent to that which can be obtained with a factorised CAT-based TTS. This result is actually very encouraging, considering that CAT-TTS can yield an expression identification rate equal to or better than that of natural speech [31].

3.7. Discussion

Some of the problems of expression similarity are due to the way in which the expressions were defined during the recording of the database. Unlike in [31] where the speaker was very carefully instructed about how to deliver each expression, here speakers were free to decide how to render them. As a result, whereas, for example, the female speaker used a “shouting at the top of her voice” kind of anger, the male speaker chose a “mounting anger” style. Since the CAT weights were defined at the sentence level, what they provide for that male angry voice is just average level ‘angriness’ somehow above that of neutral speech but not enough to make it a clear angry voice. One way to improve this would be to use CAT-weights trajectory at a sub-sentence level, e.g. syllable or words. Some oracle experiments in this direction suggest that this form of expression control can yield much more natural expressions than the static sentence-based one.

Another point that was observed was the influence of the supra-segmental models like the Global Variance. For example, one of the defining characteristics of a sad expression is a very monotonous F0. The easiest way to obtain this is by reducing the mean of the GV model for log-F0. Interestingly, this is precisely what the transformed GV models yielded: for the female input speaker $\mu^{GV}$ for sad change was reduced more than six times from 0.0391 to 0.0061, whereas for the male speaker it was halved, from 0.045 to 0.022. An advantage of the proposed framework is that those suprasegmental models can be easily integrated into it [32].

4. Conclusion

We have proposed a new technique for expression conversion which could be used to provide more flexibility to unit selection systems. The experiments have shown that the concept works. Still there is substantial work needed to improve the quality, which at the moment is basically still that of vocoded speech. For example, a better type of vocoder such as [33] could be used, which also allows for modification of aspects of the source excitation that are considered important for expression [34]. A comparison between the CAT-based transform and a full CMLLR-based transform is also necessary, but a proper training of such factorised Average Voice Model (AVM) is not trivial and would require additional technical advances. Expression conversion with non-factorised AVM models also needs to be investigated. Finally, evaluation of the system when the target expression is a complex blend of the base expressions or is allowed to change dynamically on a sub-sentence level is left for future work.
5. References


